RTE

UNIVERSITY LIBRARY,
UW-STEVENS POINT

RIVERSIDE TEXTBOOKS IN EDUCATION

EDITED BY ELLWOOD P. CUBBERLEY

DEAN EMERITUS, SCHOOL OF EDUCATION
LELAND STANFORD JUNIOR UNIVERSITY

Digitized by the Internet Archive
in 2026

https://archive.org/details/mentaltest0000fran

MENTAL TESTS

Their History, Principles & Applications

By FRANK N. FREEMAN
Professor of Educational Psychology
The University of Chicago

REVISED EDITION

HOUGHTON MIFFLIN COMPANY
BOSTON · NEW YORK · CHICAGO · DALLAS · ATLANTA · SAN FRANCISCO
The Riverside Press Cambridge

151
F87a

COPYRIGHT, 1939
BY FRANK N. FREEMAN
COPYRIGHT, 1926, BY FRANK N. FREEMAN

ALL RIGHTS RESERVED INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK OR PARTS THEREOF IN ANY FORM

The Riverside Press
CAMBRIDGE • MASSACHUSETTS
PRINTED IN THE U.S.A.

LB
1131
F83
1939

Editor's Introduction to Revised Edition

THE past quarter of a century has been a most fruitful period in the development of Education, not only as a teaching subject, but as a means for the experimental study of educational problems as well. During this time much new subject-matter for the instruction of students has been developed, and much new technique has been worked out and made applicable to the treatment of the results of investigation. It is not too much to say that the subject-matter of Education has been entirely made over during this twenty-five-year period. The new teaching material and techniques which have been evolved have been of many different types, but no aspect of this development has awakened more widespread interest, challenged the thinking of more young workers, or been more fruitful in results than the creation of tests and the application of statistical procedures to the interpretation of the results obtained.

The test movement has taken two main directions. One has been the creation of educational tests, by means of which we have been able to measure the results of the teaching process in many of its special phases; the other has been the evolution of mental tests of a number of types, by means of which we have sought to determine general intelligence and special aptitudes for training. The first aspect of the movement has been longer under way, and has by now resulted in an extensive series of educational tests for the measurement of instructional results in the different school subjects; the other has resulted in the evolution of individual and group tests for general intelligence, various types of scales, and, more recently, the application of the idea to the creation of personality and aptitude tests of a number of types. It is in this second phase of the test movement that the largest creative work is now being done.

39577

For the educational tests we have for some time had a number of teaching texts covering the field in more or less detail, but for the domain of mental tests there was no adequate presentation in a single volume until the appearance of the first edition of the present work, first published in 1926. After a dozen years of large usefulness as a textbook in the field of mental testing, the author has now thoroughly revised and in part rewritten the earlier volume so as to present the important developments that have since taken place in the subject. How thorough the revision has been the author states in detail in his preface to this edition.

The past dozen years have been important ones in the field of mental testing. In particular there has appeared a complete revision, extension, and doubling of the Stanford-Binet Scale, a large number of personality and aptitude tests and scales have been worked out and put into use, the technique of testing has been further developed and perfected, and much new evidence as to the nature of mental ability itself has been brought forth, especially under the headings of factor analysis and the study of primary abilities. The author in his revision has fully covered these more recent developments and studies and theories, thus making his text again, as it was originally, a comprehensive presentation of the field of mental testing up to the time of issue. The work of hundreds of individual investigators has been organized into a systematic treatise, and the place and work of each have been given their proper setting as parts of a great movement. The volume is accordingly again offered to teachers of college and university classes in the study of mental testing, and to individual students of the field, with confidence that it will prove even more useful as a textbook than did the earlier edition.

Ellwood P. Cubberley

Preface to Revised Eaition

THE development of mental tests during the past dozen years has necessitated a thoroughgoing revision of this book. The general plan of organization remains the same. The early chapters present an historical account of the development of mental tests; the middle chapters discuss the technique of testing and of the interpretation of the scores; and the concluding chapters discuss the application and interpretation of tests.

The historical chapters have been extended and brought down to date. This involved, for example, an extension of the description of the correlation method to include factor analysis, the account of the new revision of the Stanford Revision of the Binet scale, a discussion of aptitude tests and tests of primary abilities, and an extensive addition to the description of personality tests.

The chapters on the technique of mental tests were revised in order to include recent developments in this field. Some parts of these chapters are completely rewritten.

The organization of the latter part of the book has been modified to provide for a rearrangement of the material so as to give it a more evident relation to educational applications. The material formerly included in the chapter on "Mental Growth" has been somewhat condensed and has been placed in a new chapter on the "Basic Facts Underlying the Educational Uses of Tests." The material in the old chapter on the "Relation of Intelligence to Delinquency" has also been somewhat condensed and put in the same new chapter. The more explicit discussion of the "Educational Uses of Tests" has been separated from this material and has been placed in a separate chapter. The chapter on "The Application of Mental Tests to Voca-

tional Guidance and Selection" has been omitted because this field has become so elaborate that it warrants a more detailed discussion than can be given in a general text. The chapters on the "Interpretation of Intelligence Tests" and on "The Nature of Intelligence" have been retained but have been completely rewritten to bring them into conformity with recent evidence and with recent theories regarding the nature of ability. The discussion of these subjects has been broadened to include not only intelligence tests but tests of other forms of ability. The various theories of ability discussed in the last chapter are presented in somewhat more comprehensive and systematic fashion than was done in the earlier edition.

Throughout the book, as in the earlier edition, the attempt has been made to present fairly the facts concerning mental tests. The book has not been written exclusively from the point of view of any one of the several theories which are in existence. Where a preference for one or another point of view has been expressed, it has been done only after the various points of view have been presented as fairly as possible. The author believes that this mode of treatment in a text is preferable to a treatment of the subject from the point of view of a single theory even although the latter might give a somewhat more systematic book.

FRANK N. FREEMAN

Preface to Original Edition

THE aim of this book is to give an account of all the important types of mental tests. No book, so far as I am aware, has yet appeared which has this scope. A number of books treat of special kinds or aspects of mental tests — chiefly of intelligence tests, but none includes a description of intelligence tests, of tests of special capacities, and of nonintellectual or personality tests. All these kinds of tests are important in their practical application, and they all involve much the same principles. It seems desirable, therefore, to treat them together.

The general descriptive part of the book is organized historically. The historical approach gives a convenient method of introducing the various types of tests, and at the same time gives the best basis for the appreciation of both the value and the limitations of tests.

Throughout the book there is an emphasis on principles as contrasted with the mere surface facts concerning mental tests. The aim is to reveal the scientific problems which are involved in the design, application, and interpretation of tests, and not merely to prepare a manual for training practitioners.

Due to the recency of the development of mental tests, many of the principles which are involved in them are not very fully agreed upon, and some are not yet very clearly recognized. In such a case, where the points under discussion are still matters of debate, and where some of them are matters of individual opinion or interpretation, it seemed desirable to present various points of view, together with the evidence and the conclusions of the author. At the risk

of departing from the prevailing practice of textbook writing, I have followed this procedure. My own conviction is that this method, which encourages the reader to weigh the evidence for himself, is preferable to a more dogmatic exposition, both for the student and the reader in general. It is my hope that, on account of the wide popular interest in mental tests and their interpretation, this discussion may be of interest to laymen who may wish to inform themselves upon them, as well as to students of psychology and education.

The point of view which I have adopted upon the most widely debated issues of interpretation is in agreement with neither extreme. I have endeavored to weigh the evidence as impartially as possible, and this evidence appears to me to indicate that mental tests, particularly intelligence tests, measure native capacity in part and education or training in part.

Certain chapters are more technical than the remainder, and may be omitted by the reader who is not interested in them without seriously breaking the continuity. These chapters are those on "Technique and Theory of Mental Tests" and "How to Tabulate the Results of Tests" — Chapters IX, X, XI, and XII.

I wish to thank my colleague, Professor Karl J. Holzinger, for suggestions concerning questions of technique and interpretation. He is, however, not to be held responsible for any of the opinions expressed in the book. I wish to thank also the authors and publishers who have given permission to copy illustrations. Acknowledgment by name is made in each case.

FRANK N. FREEMAN

Contents

I. Introduction: Present Status, Meaning, and Fields of Application of Mental Tests 1

II. Early Experimentation with Tests . . 34

III. The Application of the Correlation Method 60

IV. Age Scales: The Binet Scales and their Revisions 85

V. The Early Development of Point Scales . 108

VI. Survey of Point Scales 141

VII. Tests for the Analysis of Mental Capacity 169

VIII. Tests of Personality Traits 205

IX. Technique and Theory of Mental Tests: I. subject-matter of tests and related problems 257

X. Technique and Theory of Mental Tests: II. problems relating to the selection and organization of the items of a test . . . 260

XI. Technique and Theory of Mental Tests: III. problems relating to scores and norms 277

XII. How to Tabulate the Results of Tests . 322

XIII. Basic Facts Underlying the Educational Uses of Tests 345

XIV. The Educational Uses of Tests 370

XV. Interpretation of Mental Tests . . . 394

XVI. The Nature of Ability 431

Index 445

MENTAL TESTS

Chapter I

INTRODUCTION: PRESENT STATUS, MEANING, AND FIELDS OF APPLICATION OF MENTAL TESTS

MENTAL tests are of recent origin. They grew out of the study of individual differences in the psychological laboratory. The study of individual differences, in turn, grew out of the experimentation which had for its aim the discovery of general principles or general laws concerning human behavior. At the beginning of this experimentation the individual variations which occurred in the course of experiments were regarded either as errors or as negligible quantities. After a time psychologists recognized that these differences were real and that they deserved to be studied directly. The study of these individual differences for themselves began about forty-five years ago. This was some fifteen years after the founding of the first important psychological laboratory by Wilhelm Wundt.

The scientific interest in individual differences and their measurement began to develop about 1890. For ten or fifteen years tests were tried out in the psychological laboratories of the universities. The educational interest in tests may be said to have begun about 1905. The development of practical tests for use in schools, therefore, began about thirty years ago. The interest in tests before this time was fostered largely by professors of psychology, and their experiments were carried on largely with the college students.

These experiments had very little immediate application to educational problems. They laid the foundation, however, for the development of tests which could be used for the practical differentiation of the pupils in the school.

The beginning of the development of practical tests is represented in the work of Binet. Binet, like other psychologists, had been experimenting with tests during the decade 1890–1900, but his labors during this early period had been of little more practical value than those of other experimenters. During the first decade of the present century, however, he succeeded in developing the scale which overcame the shortcomings of the earlier tests and which proved to be of immense practical value. The outstanding characteristic of this decade was the development of the individual scale of the type of the well-known Binet-Simon scale. This test was applied particularly to the discovery of backward, subnormal, and feeble-minded children, in order that they might be assigned to special classes.

While the dominant interest during this period was in groups of tests which are represented by the age scales of the type of the Binet scale, there was also a considerable amount of experimentation going on with single tests by means of the method of correlation statistics. While the study of these single tests by means of the correlation method did not at the beginning prove very fruitful, it led, during the following decade, to types of experimentation and the development of types of tests which proved to have still wider application than the individual tests of the age-scale type. These later tests are the prevalent group point scales. The past two decades saw an enormous development of these group tests, which can be applied conveniently to children on a large scale. The most important factor in the large-scale development of these tests was, of course, the World War, which was the occasion of the pro-

duction of the army scales, and of scales patterned after them for use in the schoolroom. Coincident with the extension of tests came the shifting of interest from the backward or feeble-minded child to the normal child, and particularly to the child of unusual ability.

During the past few years an enormous number of group tests has been given. Over 1,700,000 men were given the Army Alpha Test. Following the War, a committee of psychologists who had been concerned with the development of the army tests formulated the National Intelligence Test. Within less than a year after this test was issued, over 575,000 copies were sold. During the year 1922–23, 800,000 copies of this test were distributed. During the same year one firm which deals particularly in mental tests had sold over 2,500,000 intelligence tests. There are, at the present time, over forty well-known group tests of intelligence on the market which are designed for use in the schools. They are adapted to stages of development ranging from the kindergarten to the university. They are used not only in the schools, but also in the industries and in the courts. The terms in which mental abilities are described have become incorporated into popular language. The possibility of measuring an individual's intelligence by a short and simple test has captured the imagination of school people and of the general public.

1. A sample intelligence test

In order that the reader may at the outset of the discussion make direct acquaintance with this one type of mental test — the intelligence test — he is here given an opportunity to examine or to take an abbreviated form of such a test. The purpose of putting the test in at this point is to give the reader an idea of what we are talking about when we discuss mental tests. He should, of course, not draw con-

clusions as to what the test measures or as to the meaning of the score until he has read the later chapters.

The following test was designed for group administration. It is graded to suit the capacity of high-school seniors and college freshmen. The original from which the abbreviated form was made up is *Test IV, Psychological Examination,* by L. L. Thurstone.[1]

The directions should be followed faithfully.

Directions

This is a test to see how quickly and accurately you can think. The result of the test will be used by your advisers in order that they may know more about your abilities.

On the inside pages there are 56 short problems. In each case you are told exactly what to do. Notice the instructions carefully. You may use the margin for figuring.

If you come to a problem that you do not understand, go to the next problem.

Take ten minutes. Solve as many problems as you can in the time allowed.

Solve the problems in order given. Do not skip about on the page.

The Test

1. Underline the correct answer.

   London is in **England** **Australia** **Brazil** **Spain**

   The correct word is **England.** Underline that word.

2. Underline the correct answer.

   Boston is in **Connecticut** **Rhode Island** **Maine** **Massachusetts**

3. Underline the correct answer.

   Diamonds are obtained from **mines** **reefs** **elephants** **oysters**

[1] This test is published by C. H. Stoelting Company and is used by permission of the author.

4. Underline two words that have the same relation as **locomotive** and **train.**

   station horse hub baggage buggy

   Underline **horse** and **buggy** because the horse pulls the buggy and the locomotive pulls the train.

5. Underline two words that have the same relation as **good** and **bad.**

   taste sweet conduct sour polite

   Underline **sweet** and **sour** because they are opposite in meaning, just as **good** and **bad** are opposite in meaning.

6. Underline two words that have the same relation as **ear** and **hear.**

   eye hair blue see eyebrow

7. Underline two words that have the same relation as **palace** and **king.**

   hut peasant barn farm city

8. Make a perfect sentence. Write one word on a blank.

   There are .....days in a week.

   Write the word **seven** in the blank.

9. Make a perfect sentence. One word on a blank.

   The boy will .....his hand if .....plays with fire.

10. If the following conclusion is true, underline **true**; if it is false, underline **false.**

    Brown is shorter than Smith. Jones is shorter than Brown. Therefore Jones is shorter than Smith.

    True False (Underline one)

11. **Don't put all your eggs in one basket.**

    Check two of the following statements with nearly the same meaning as the above proverb:

....The mouse that has but one hole is soon caught.
....Catch the bear before you sell his skin.
....The proof of the pudding is the eating.
....Put not all your crocks on one shelf.

Check the first and fourth statements.

12. Tall oaks from little acorns grow.

Check two of the following statements with the same meaning as the above proverb:

....No grass grows on a beaten road.
....Large streams from little fountains flow.
....The exception proves the rule.
....Great ends from little beginnings.

13. Write the two numbers that should come next.

2 4 6 8 10 12 ... ...

The two numbers are 14 and 16.

14. Write the two numbers that should come next.

2 2 3 3 4 4 ... ...

15. Write the two numbers that should come next.

1 7 2 7 3 7 ... ...

The two numbers are 4 and 7.

16. Write the next two numbers.

1 4 7 10 13 16 ... ...

17. Underline the correct answer.

Arthur Brisbane is famous as a **newspaper man** **comic artist** **athlete** **actor.**

18. Underline two words with the same relation as **egg** and **bird.**

crack **seed** **plant** **grow** **nest**

19. Make a perfect sentence. One word on a blank.

The poor.....is hungry because.....
has.....nothing to.....

20. John's birthday is after Harry's, and Harry's birthday is after Tom's. Therefore Tom's birthday is before John's.

True **False** (Underline one)

21. **"Every one of us, whatever our speculative opinions, knows better than he practices, and recognizes a better law than he obeys."** (Froude.)

Check two of the following statements with the same meaning as the quotation above:

....To know right is to do the right.
....Our speculative opinions determine our actions.
....Our deeds fall short of the actions we approve.
....Our ideas are in advance of our every day behavior.

22. Write two numbers that should come next.

14 16 18 20 22 24 ... ...

23. Underline the correct answer.

Yale University is at **Annapolis Ithaca Cambridge New Haven**

24. Underline two words with the same relation as **foot** and **man.**

hoof leather shoe cow leg

25. Underline the correct answer.

"The makings of a nation" is an ad. of a **tobacco flour beer health food**

26. Underline two words with the same relation as **wash** and **face.**

sweep broom floor straw clean

27. Make a perfect sentence. One word on a blank.

It is very.....to become.....acquainted .....persons who.....timid.

28. Since all metals are elements, the most rare of all the metals must be the most rare of all the elements.

True False (Underline one)

29. A small leak will sink a ship.

Check two of the following statements with the same meaning as the above proverb:

....Untempted virtue is easily retained.
....A spark may start a great fire.
....When the cat is away the mice will play.
....Reputation may be ruined by a word.

30. Write the two numbers that should come next.

2 3 5 8 12 17 ... ...

31. Underline the correct answer.

Dioxygen is a disinfectant food product patent medicine tooth paste

32. Underline the two words with the same relation as skating and winter.

swimming diving floating hole summer

33. Underline the correct answer.

The Corona is a kind of phonograph multigraph adding machine typewriter

34. Underline two words with the same relation as able and unable.

muscle exercise strong ax weak

35. Make a perfect sentence. One word on a blank.

The dog.....a useful.....because..... his intelligence and faithfulness.

36. All the members of the Civic Club are members of the University Club; Smith is not a member of the University Club; therefore he is not a member of the Civic Club.

True **False** (Underline one)

37. **"Equality is the life of conversation; and he is as much out who assumes to himself any part above another, as he who considers himself below the rest of society."** (Steele.)

Check two of the following statements with the same meaning as the above quotation.

....One should assume himself below those with whom he converses.
....One should not consider himself on a different level from those with whom he converses.
....One must talk or be talked to, there is no middle ground.
....Conversation should be democratic.

38. Write the two numbers that should come next.

28 31 33 36 38 41 ... ...

39. Underline the correct answer.

The Delco System is used in **plumbing** **filing** **ignition** **cataloguing**

40. Underline two words that have the same relation as **telephone** and **hear.**

shout **telegraph** **spyglass** **distance** **see**

41. Underline the correct answer.

Darwin was most famous in **literature** **science** **war** **politics**

42. Underline two words with the same relation as **reward** and **hero**.

God everlasting punish pain traitor

43. Make a perfect sentence. One word on a blank.

A home is.....merely a place.....one .....live comfortably.

44. All double-convex lenses magnify; plano-convex lenses are not double-convex; therefore plano-convex lenses do not magnify.

True **False** (Underline one)

45. **Familiarity breeds contempt.**

Check two of the following statements with the same meaning as the above proverb.

....Every bird likes its own nest best.
....Sweets grown common lose their dear delight.
....Birds of a feather flock together.
....No man is a hero to his valet.

46. Write the two numbers that should come next.

42 41 37 36 32 31 ... ...

47. Underline the correct answer.

The battle of Lexington was fought in 1620 1775 1812 1864

48. Underline two words with the same relation as **floor-walker** and **store**.

policeman fire street conductor wagon

49. Underline the correct answer.

The Overland car is made in Toledo Flint Buffalo Detroit

50. Underline two words with the same relation as **table** and **wood.**

stove **bottle** **paper** **iron** **cork**

51. Make a perfect sentence. One word on a blank.

The.....is always shining.....storm clouds sometimes.....it.....us.

52. No Athenians could have been Helots, for all Helots were slaves, and all Athenians were free men.

True **False** (Underline one)

53. **"No great genius was ever without some mixture of madness, nor can anything grand or superior to the voice of common mortals be spoken except by the agitated soul."** (Aristotle.)

Check two of the following statements with the same meaning as the above quotation.

....Genius is essentially hard work and persistence.
....Contented and serene characters are the ones that produce works of genius.
....Genius and insanity have certain elements in common.
....Strokes of genius are likely to come after times of great disturbance or stress for the individual.

54. Write the two numbers that should come next.

15 18 24 33 45 60 ... ...

55. Underline the correct answer.

Plymouth Rock is a kind of **horse** **cattle** **granite** **fowl**

56. Underline two words with the same relation as **Japanese** and **Japan.**

Dutch **Russia** **Holland** **Siberia** **Spanish**

57. Underline the correct answer.

Rio de Janeiro is a city of **Spain** **Argentina** **Portugal** **Brazil**

58. Underline two words with the same relation as **quarrel** and **enemy.**

policeman **agreeable** **foe** **agree** **friend**

59. Make a perfect sentence. One word on a blank.

.....things are.....satisfying to an ordinary.....than congenial friends.

60. The recent panic occurred just after the President announced his policy regarding corporations in interstate commerce; therefore the President is to blame for the panic.

True **False** (Underline)

61. **Rome was not built in a day.**

Check two of the statements with the same meaning as the above proverb.

....To climb steep hills requires slow pace.
....When in Rome, do as the Romans do.
....The result tests the work.
....Napoleon himself was once a crying babe.

62. Write the two numbers that should come next.

19 **21** **23** **18** **20** **22** ... ...

63. Underline the correct answer.

The spark plug belongs in the **carburetor** **manifold** **crank case** **cylinder head**

64. Underline two words with the same relation as **eat** and **fat.**

food **starve** **thin** **bread** **thirsty**

The reader may now check the number of items he has marked correctly by referring to the key in the appendix to

the chapter on page 30. The score is found by counting the number of correct items.

The median score obtained by high-school seniors or college freshmen when the test is taken under ordinary conditions is about 30.

The first types of mental tests that were developed to such a form as to be of practical use in the schools were the individual and group intelligence tests which have been described. More recently two other general types of tests have been worked out, tests of special abilities and tests of personality.

Tests of special abilities represent, in a measure, a return to the kind of tests which were experimented with in the nineties, before Binet devised the test of intelligence. Some of the abilities tested are more complex than those of the earlier period. The present tests aim to measure mechanical ability, musical ability, artistic ability, mathematical ability, and language ability. Others are not very different from those tested by the earlier psychologists, for example, memory, perception, imagination. Simple sensory and motor capacities, which were prominent in the early tests, are still used for special purposes.

The most prominent development since the World War has been in the testing of personality. This is a loose term covering tests of emotions, attitudes, interests, social behavior, ethical judgment, conduct, nervous stability, and temperament. They are illustrated in Chapter VIII. Many of these tests have been tried out experimentally in the schools, but their interpretation and practical application are not yet as clear as are those of intelligence tests or even of tests of special ability.

2. Reliability and meaning of mental tests

Following the great expansion of testing during and after the World War a lively discussion ensued concerning the

reliability and significance of intelligence tests. This discussion was carried on chiefly by lay writers, or at least by observers who did not have specialized training in psychology or an intimate knowledge of mental tests. Popular opinion goes to greater extremes concerning mental tests than does the opinion of psychologists.[1] It is marked both by more implicit confidence in them and more extreme skepticism.

Similarly, popular opinion shows the sharpest fluctuation from one period to another. Before the World War, the average intelligent layman probably had little confidence in the value or the use of mental tests. After the War, he believed that psychologists had devised a simple and relatively perfect method of measuring intelligence.

That popular opinion should be subject to more extreme fluctuation than the opinion of psychologists is easily understood. The psychologist realizes that mental tests are the product of a long period of experimentation. He knows that our present-day tests have been developed from earlier forms of tests, that they constitute an improvement over the earlier forms, but that they are subject in some degree to the limitations of the first attempts. He knows what the difficulties were which confronted the earlier experimenters, and the methods which were adopted to overcome these difficulties. He knows, furthermore, that these difficulties have been only partly met.

The layman, on the other hand, has been accustomed to think of mental tests as something absolutely new. He regards them as an invention. He believes that psychologists have made a clear analysis or classification of human abilities, and that they pretend to have de-

[1] For the agreement among psychologists see Frank N. Freeman, "A Referendum of Psychologists," *Century*, 107 (December, 1923), 237–45.

vised methods by which these abilities may be perfectly measured.

This view seems to leave the way open for only one of two extreme conclusions. One may either regard mental tests as wholly successful, or he may entirely reject them. This view of the tests seems to furnish no opportunity for an intermediate view. There is no basis upon which one may form an opinion of the limitations of the tests and of the range of the problems to which they are adapted.

A correct knowledge of the nature and development of mental tests shows the absurdity of the fundamental assumption which underlies the popular view. Mental tests are not absolutely new devices. They are not magical instruments for the discrimination and measurement of mental capacities. Their fundamental characteristics are the same as those of the ordinary examination with which we have been familiar so long.

The methods of the examination have, of course, been very greatly refined. Students of mental tests have discovered what will work and what will not work. They have devised methods of organizing tests so that they will measure the abilities which are the most significant factors in general human behavior. They have discovered methods of making the tests so widely applicable that broad comparisons can be made by means of them.

Mental tests enable us to secure with comparative ease a more widely comparable measure than could be secured by any other means. Their accuracy, furthermore, is at least comparable to that of the best methods which exist beside them. In addition to this, they enable us to make a type of analysis of mental abilities which we cannot make as satisfactorily by any other means. All these advantages of mental tests must be granted.

At the same time, it must be recognized that enthusiastic

advocates of mental tests sometimes give excuse for the undue enthusiasms of the layman. While we recognize the advantages which they offer, we must not exaggerate the accuracy of the measures which they yield. We must recognize that the ratings of human capacity which they enable us to make are correct only within certain limits of error.

We must recognize further that the nature of the capacities which they measure are known to us only in a rough way. The interrelationships between the abilities which are measured by various tests are often very surprising. They indicate that abilities which we would not expect to be closely related do, as a matter of fact, correspond very closely, and abilities which we are accustomed to believe closely associated are really comparatively independent of one another.

The advancement of the technique of mental tests will be furthered most by a sane recognition of both the advantages and the limitations of the present tests. We may be grateful for that which they furnish us, without exaggerating what they have to offer. An unwarranted satisfaction with present tests will prove a hindrance to the experimentation which is necessary for the development of tests of greater accuracy, greater range, and of more analytical power than our present tests possess.

Our present tests are most successful as measures of the composite of mental abilities which is sometimes called intelligence. The chief problem in the development of mental tests at the present time is the need of more precise definition of the abilities or traits which are to be measured. In the past, abilities or traits have been defined chiefly in terms of the performance on tests, and the tests have required composites of abilities or forms of behavior. The value of the tests has been judged in terms of their work-

ability in practical situations. Thus, intelligence is what is measured by intelligence tests and the intelligence tests are useful because they enable us to predict performance in school and in some other situations.

For many years certain psychologists, particularly Spearman, have not been satisfied with this kind of definition. They have tried to obtain a more precise analysis and definition by means of a statistical study of the scores of tests and the development of new tests which will give a clearer separation of abilities.

In recent years this statistical investigation, called factor analysis, has been vigorously pursued. It has led to somewhat divergent theories, and its bearing on test construction is not yet clear. That it will have an important bearing on the future development of mental tests seems evident.

3. Definition and classification of tests

A test may be fundamentally distinguished from a descriptive account of a mental function. Both the descriptive account and the mental test involve the use of accurate experimental technique, but the aims of the two forms of procedure are different. The aim of the descriptive account is to determine how a mental function operates, how it develops, what its causes and effects are. We may be interested, for example, in breaking up a perception into its constituent sensations. We find that what looks at first glance like a simple experience is really a complex one. We may be interested in learning the causes of peculiar experiences, such as optical illusions. In doing this, we again break up or analyze the experience into its parts. We may be interested in studying the growth or development of a mental ability. Some studies are concerned with the learning process. They trace the effect of practice upon skill or

knowledge. A somewhat similar study is the investigation of the development of the mental capacity in the child as he grows from babyhood to maturity, or the somewhat parallel development of the mental life of animals from the lowest organisms to the higher mammals.

In contrast to these types of study and to these aims, the mental test seeks to measure the strength, precision, or effectiveness of the present operation of any mental activity. It does not aim to determine how the activity was developed or in what it originated, or, necessarily, the elements of which it is composed. It takes the ability as it exists at the present time, and attempts to set up means of estimating its degree. In the case of sensation, for example, a test aims not to analyze the sensation, but to determine the ability of the individual to discriminate between sensory stimuli which differ by a slight degree. In the case of learning, the test seeks not to discover how one learns, but to discover the rapidity or accuracy with which one person can learn in comparison with other persons.

Mental tests, again, may be distinguished from educational tests. The aim of both alike is to measure the present efficiency of the individual in certain specific respects. They differ, however, in this: whereas the educational tests seek to measure the products of training, and indirectly to determine the efficiency of the training which the individual has received, the mental tests aim to measure the original capacity which the individual had for the acquirement of skill or knowledge or ability. The extent to which educational and mental tests are able to meet the demands of these contrasted aims is a matter to which we shall have to give attention later in the discussion.

We may note in passing that mental tests and educational tests have often been studied in relation to one another, in order that it might be determined what the ratio is be-

tween the inherent capacity of an individual or group of individuals and the actual achievement which they have made.

Tests, whether they be mental tests or educational tests, are both relative. That is, the score which results from the application of the test has significance only by comparison with scores which are made by other individuals. The score serves as a comparatively exact numerical method of indicating the rank of the individual in a group in which he may be placed or with which he may be compared. The absolute score which the individual makes, taken by itself, has, therefore, no significance. The method by which is expressed the relationship between the score of an individual and the scores of the group constitutes one of the important phases of the technique of mental testing.

Mental tests are of various kinds, according as they aim to measure general or special capacities. The tests which are most widely used by educators at the present time are general tests, usually called *general intelligence tests*, or sometimes *mental alertness tests*. There are, however, a considerable number of tests in use which aim to measure not all-round or general intellectual capacity, but which, on the other hand, aim to measure some special capacity or set of capacities. An example of such tests is the collection of tests of musical ability designed by Seashore. A good many tests of special capacity have been used in vocational guidance. The purpose of this book will be to include a treatment of the special tests as well as the general ones.

We sometimes speak of tests as though they measured intellectual capacity directly. This, of course, is not true. What they measure is the manifestation of capacity in action or in behavior. Intellectual capacity is not something which can be seen, felt, heard, or measured in any direct fashion. We assume in mental tests that the behavior of the individ-

ual expresses or represents the maximum of which he is capable.

Behavior, however, is always conditioned, not only by capacity, but also by previous experience or training. A person is able to play a piano or to write on a typewriter not only because he has the capacity for learning to use these instruments, but also because he has gone through a course of training. These are specific but rather extreme cases. To take a more widely applicable and general case, a person is able to use language because he has come in contact with language and has acquired the ability to pronounce words and a knowledge of the meaning of words. Somewhat more generally still, a person has learned to distinguish a color because he has met with different colors. He has learned to distinguish the pitch of tones because he has met with tones of different pitches.

In all of these cases, or in any case which might be mentioned, the capacity which the individual has to start with is combined with the results of his training to make up the ability which he possesses at the present moment or at the moment of being tested. If training is thus always present as a factor in determining present ability, how is it possible to distinguish native capacity from the results of training? To put the question in a somewhat more specific way, how can we determine that the differences between individuals are due to differences in their capacity, rather than to differences in the training which they have received?

Mental tests are so designed as to meet this difficulty, so far as possible, in the following manner: The particular activities which are demanded of the individual being tested are selected from among those which are common to the experience of all the persons who are to be compared. Or, to put it in another way, it is assumed that training or experience in the activities which are being tested are equal,

or as nearly as possible equal, among all the individuals. For this reason, typewriting or piano-playing, for example, would not be taken as the subject of a mental test for motor dexterity or manual capacity. These particular complex types of behavior require special practice for their mastery. If one wished to test capacity in learning of this type, it would be necessary to devise some activity of a similar nature to these, but one which had never been practiced by any of the individuals who were to be tested. The measurement of the ability of the individual to distinguish the pitch of two tones, or the shade of two colors, is, in contrast to typewriting or piano-playing, a suitable test for native capacity, because all the individuals who are likely to be tested have had sufficient practice in doing these things to bring their capacities up to something near their maximum. This is partly due to the fact that everybody has had practice in these things, and partly to the fact that they are not so susceptible to practice as the more complex activities of piano-playing or typewriting.

It may, of course, be questioned whether it is ever possible to find activities in which the individuals who are to be compared have had equal opportunity for training, or in which training is a negligible factor. We are faced here with a dilemma. If we choose to measure abilities in which training has little effect, we find that our measurements have very little general significance. If, on the other hand, we select abilities which are complex in their nature, and which are therefore of general significance, we find it difficult to secure activities in which previous experience is not an important factor. This is one of the problems concerning which a knowledge of the development of tests and their technique will enable us to be duly critical and on our guard.

We may summarize this descriptive account of mental tests in the following definition: "*Mental tests are instru-*

ments for the measurement of individual abilities or types of behavior, with maximum emphasis on differences due to original nature rather than to training or environment."

Mental tests are instruments of measurement and not means of making guesses or estimates. They are therefore to be distinguished from methods of rating individual abilities by means of rating scales. They issue in numerical scores which can be manipulated by mathematical processes and combined or compared with other numerical scores. The method by which these scores are obtained may not, of course, be valid, but it is of advantage that the results of the tests be thus expressed in quantities which are subject to mathematical formulation.

The significance of the measurements which are made by means of mental tests grows out of the fact that the tests are standardized. This standardization concerns the materials which go into the tests, the method of procedure in giving the tests and in scoring the results, and the norms with which the scores of individuals are compared. Standardization, in brief, means that all of these matters are worked out by actual trial. The materials and methods of procedure are not invented by some psychologist in the seclusion of his study, but they are arrived at after the tests have been given to large numbers of children and after the procedure has been discovered which proves to be successful.

The measurement is relative, as already noted, because the score which any individual makes is to be interpreted in comparison with the scores which are made by other individuals. This comparison, in the case of tests which aim to measure native capacities rather than the results of training, is completely satisfactory only in groups which have had approximately equal opportunity, or with reference to mental functions in which differences of training or opportunity have a very slight effect. It is not possible to deter-

mine with exactness just how far a test score may be due to training or to native capacity. This constitutes one of the large problems, both in the organization and the interpretation of tests, which we shall have to discuss at greater length in a later chapter.

4. The uses of mental tests

We may anticipate the more detailed discussion of the applications of mental tests, in the later chapters, by giving a brief summary or survey of their uses, as a means of introducing us to the fuller description of the characteristics and organization of mental tests themselves. Some of these uses are practical and some theoretical. We are more immediately concerned with the practical uses than with the theoretical uses. It may ultimately turn out, however, that the theoretical interpretations of the results of mental tests may have a more far-reaching practical effect than their immediate practical application.

The first use to which tests are put is the classification of pupils in school according to ability. It has long been recognized that pupils differ very widely in their capacity to do school work. This recognition first became clear with reference to feeble-minded or backward children. The fact of backwardness was forced upon the attention of school authorities by the large amount of retardation and elimination which it produced. After tests had been made of all children, and the distribution of pupils' abilities had been tabulated, however, it was discovered that as many pupils possessed extremely high as extremely low ability. In fact, the evidence has gone to show that the pupils are distributed with reference to their intellectual ability in the same fashion above the median as below it. The problem of classification, therefore, is a much broader one than is represented in the selection of a few pupils for special instruc-

tion because of their mental deficiency. It is only within the past two decades or so, however, that this has been clearly recognized as a problem.

The classification of pupils according to abilities may be either *vertical* or *horizontal*. By vertical classification is meant the arrangement of pupils at successive levels of attainment or advancement through the school. According to this method the pupils who are able to do a certain grade of work are placed in a particular school grade. Those who are able to do a higher grade of work are placed in a higher grade. Promotion, according to this type of classification, is based on mental capacity alone without regard to any other factor. This may be called vertical classification.

Horizontal classification, on the other hand, takes all the pupils at a given stage in school advancement, who may be widely diverse in their ability to do work at that stage, and groups them according to their ability. They are grouped into horizontal divisions according to ability, and their grouping in vertical divisions is based simply upon their age. If this type of classification alone is carried out, it does not affect promotion through the school, but does affect the difficulty of work or the quality of work which a pupil does in each successive grade. Whether vertical grouping or horizontal grouping is the better, or whether some combination of the two methods is better than either alone, is a problem for later, more detailed consideration.

A second general use of tests in the school is to serve as a means of diagnosis of the capacity of pupils who present problems in adjustment because of their failure to do successfully part or all of the work of the school. The needs of individual diagnosis and the subsequent treatment are not entirely met by the general classification which has already been spoken of. The failure of the pupil may not be due to inherent incapacity. His classification in a low group,

therefore, does not solve the problem. The problem which is presented may be to find a means of awakening the pupil to the realization of his capacity. The poor work may be due to a variety of causes. Mental tests contribute to the solution of the problem by indicating the extent to which the poor work is caused by incapacity. While the results of the test are not absolutely conclusive in regard to the pupil's capacity, they do, at least, indicate a line of experimental treatment which may result in the solution of the difficulty.

A third use of tests consists of educational guidance. Some pupils may, by reason of the degree of general intellectual ability they possess, or because of the type of their capacity, be better adapted to some courses of study than to others. Again, the length of time a pupil can profitably remain in school may be determined by his native capacity. Mental tests serve, therefore, as partial means of estimating the kind or extent of work for which the pupil's capacity suits him. Educational guidance takes the form of advice in the selection of courses or of the larger groups of courses or curricula, or of advice concerning the desirability of remaining in school or college or of going to work.

Educational guidance naturally leads into and prepares the way for vocational guidance. The selection of the types of work which the pupil takes in school looks forward to the type of vocation which he shall pursue. Vocational guidance takes the matter up at the point where educational guidance leaves it, and attempts in a more specific way to aid the pupil in the choice of a vocation.

Tests for vocational guidance may be tests of general ability, or tests of special ability. The use of tests of general capacity is based upon the assumption that various vocations require for their successful pursuit different degrees of intellectual capacity. If this assumption is correct, it is possible within certain limits, by means of tests, to deter-

mine those groups of vocations which a person can expect to pursue successfully. It would not, of course, determine which among a group demanding equal ability he should choose. Tests of specialized ability aim more specifically to determine whether one can meet the requirements of a particular vocation, other than the requirement of general capacity. These tests are, for the most part, applicable to specialized jobs in industry or to specialized phases of the work of a more general vocation. They may consist of single tests of simple mental functions or of groups of tests which measure a variety of functions, all of which are required in the vocation, or of tests which measure a complex activity involving a variety of separate capacities. Examples of all of these types of tests are in existence and have been tried in vocational guidance.

The application of mental tests for vocational selection, as distinguished from guidance, consists in their use in selection, transfer, or promotion of employees. Tests are here used not to determine what vocation an individual should enter, but to determine whether or not individuals who may wish to enter a particular vocation meet the conditions of that one particular vocation. This concerns the selection of employees. Transfer involves somewhat similar principles. Employees who are in one department of an organization, and who may not be suited to the work of this department, may be more capable of performing the work of some other department. Tests may be given to them to determine this fact. Promotion may be governed in part by capacity as measured by tests.

The use of tests for vocational guidance or for the selection of employees depends, of course, on whether or not vocations do differ largely in their demands, and whether individuals differ in the possession of capacities to meet these demands. Upon this question there may be a diverg-

ence of opinion. The complete solution of the problem rests largely with future experimentation.

Again, mental tests have been applied to delinquents. Their purpose is to assist in fixing the degree of responsibility and in indicating the kind of treatment which should be given. Tests have been widely used in dealing with juvenile delinquents and somewhat less extensively with adult offenders. Widely different views were formerly held with reference to the significance and the interpretation of the results. On the one hand, the opinion was rather general that mental incapacity is responsible for a very large share of crime. On the other hand, it was believed that while mental deficiency is responsible for certain cases of crime, and particularly for certain types of crime, it is, on the whole, a relatively minor cause. The result of these tests will be reviewed more at length in a special chapter.

The final practical use of tests of ability which will be mentioned here is the measurement of the efficiency of educational units. By efficiency is meant the relationship between achievement and capacity. In speaking of the efficiency of an individual, we ask ourselves not merely what capacity he has, but whether he realizes his capacity in productive activity. If we assume that mental tests measure native capacity, and educational tests measure the result of training, it follows that the relationship between the scores on educational tests and scores on mental tests will represent the efficiency of the individual or of the group. This relationship has been expressed in the form of a ratio which is called the *achievement quotient*, or the *accomplishment ratio*. The validity of the achievement quotient, of course, depends upon the clearness of the distinction between the measurement of native capacity and of training, and upon the accuracy of the measures of both of these factors.

A somewhat more theoretical problem is the determina-

tion of the character of the mental growth of children. The multitude of scores of children of various ages which have been gathered from the application of various mental tests yields a mass of material which gives a basis for more valid estimates of the character of intellectual growth than we have previously possessed. It is true that these measures have usually been limited in one respect. They have been made upon different children of different ages. They have not, that is, given successive measures of the capacity of the same child at successive periods in its growth. They therefore give us only a mass picture of the general characteristics of growth, and do not enable us to determine what the fluctuations in the case of individuals are. Beginnings are being made in the successive testing of individual children, so that we may, in time, possess a more accurate picture of intellectual development.

One of the long-standing problems with reference to human capacity concerns the relative effect of the factors of heredity and environment. The question is, do differences between individuals depend largely on differences in their inherited mental traits, or are they the product of differences in training and the less definable features of the general mental and physical environment. As we have already seen, the relative effect of heredity and environment is in reality a problem in the interpretation of the test scores themselves. It might seem, therefore, as though test scores could not be used as means of determining the relative share which native capacity has in any other form of achievement. The problem presents difficulties, and we are only at the beginning of its solution, but there are methods, as we shall see in a later chapter, by which we may at least make some advance toward its solution.

Finally, mental tests furnish means of studying the interrelationship of mental traits and of investigating

mental types. It is, of course, a moot question whether mental types exist. By types is meant constellations or groups of abilities which are frequently found to exist in conjunction with one another. One view is that there do not exist such types, but that the various mental abilities are just as likely to be associated in one way as in another. The study of the correlation of the scores of mental tests will ultimately enable us to determine whether such types exist. The difficulty with the interpretation at the present time is that the tests themselves do not measure very clearly definable characteristics. There is a large overlapping in the functions which are measured by the various tests. This is one of the problems for the future which rests, in part, upon the development of a somewhat new type of test itself.

The applications of tests of personality have not been worked out so definitely as have the applications of tests of ability. This may be due either to the fact that they are newer or to their inherent limitations. Time will tell.

As distinguished from tests of ability, tests of personality have not been extensively used as the basis for routine administrative practices, such as the classification of pupils, or educational or vocational guidance. It does not seem possible, thus far, to put pupils into groups on the basis of their responses to tests of personality which demand distinctive educational treatment.

In the meantime, much research is being carried on by means of personality tests to determine how far the characteristics revealed by them influence or are associated with academic achievement, conduct, social adjustment, or vocational success. Such research may ultimately lead to broader forms of application than are as yet feasible.

The chief application of personality tests up to the present is found in clinical examination and diagnosis. Skilled psychologists or psychiatrists are able to use them in con-

junction with tests of ability, life histories and interviews to gain an insight into the constellation of factors which determine the individual's behavior. They are therefore valuable tools in the hands of experts but are not suitable for use by those without special training.

We have now reviewed briefly the topics or questions which will be discussed in a more specialized fashion throughout the different chapters of the book. We shall first review at somewhat greater length the historical development of tests, in order that we may bring out the contrast between the earlier less successful attempts and the later more successful ones. We shall then pass in review the different kinds of tests, examining in some detail the most prominent tests which are of practical importance in the school. Finally, we shall consider the uses of tests, particularly those which have important applications in education. Throughout the discussion we shall emphasize particularly those aspects of tests which are of practical importance.

CORRECT ANSWERS TO THE ITEMS OF THE TEST ON PAGE 4

2. Massachusetts
3. mines
6. eye, see
7. hut, peasant
9. burn, he
10. true
12. numbers 2 and 4
16. 19, 22
17. newspaper man
18. seed, plant
19. man, he, had, eat (or equivalent)
20. true
21. third and fourth
22. 26, 28
23. New Haven
24. hoof, cow
25. tobacco
26. sweep, floor
27. difficult, well, with, are
28. false
29. second and fourth
30. 23, 30
31. disinfectant
32. swimming, summer
33. typewriter
34. strong, weak
35. is, animal, of
36. true
37. second and fourth
38. 43, 46
39. ignition

40. spyglass, see
41. science
42. punish, traitor
43. not, where, can
44. false
45. second and fourth
46. 27, 26
47. 1775
48. policeman, street
49. Toledo
50. stove, iron
51. sun, but, hide, from
52. true
53. third and fourth
54. 78, 99
55. fowl
56. Dutch, Holland
57. Brazil
58. agree, friend
59. few, more, person
60. false
61. first and fourth
62. 17, 19
63. cylinder head
64. starve, thin

SELECTED LIST OF GENERAL BOOKS ON MENTAL TESTS AND THEIR APPLICATION

Bingham, W. V. *Aptitudes and Aptitude Testing.* New York: Harper & Bros., 1937.
A general book, written in simple style, describing the practical use of aptitude testing.

Bronner, Augusta F., Healy, William, Lowe, Gladys M., and Schimberg, Myra E. *A Manual of Individual Mental Tests and Testing.* Boston: Little, Brown & Co., 1927.
A detailed description of tests and their administration.

Dearborn, Walter Fenno. *Intelligence Tests.* Boston: Houghton Mifflin Co., 1928.
A critical account of the nature of ability and of the methods of testing it.

Dickson, Virgil E. *Mental Tests and the Classroom Teacher.* Yonkers-on-Hudson, New York: World Book Co., 1923.
A detailed account of the use of tests in the school, with constant reference to the author's experience in Oakland, California.

Garrett, Henry E., and Schneck, Matthew R. *Psychological Tests, Methods, and Results.* New York: Harper & Bros., 1933.
A description of tests classified according to psychological functions.

Hildreth, Gertrude H. *A Bibliography of Mental Tests and Rating Scales.* New York: Psychological Corporation, 1933.

Hines, Harlan Cameron. *Measuring Intelligence.* Boston: Houghton Mifflin Co., 1923.
A brief, somewhat theoretical account of intelligence tests and their interpretation.

Hollingworth, H. L. *Vocational Psychology.* New York: D. Appleton & Co., 1923.
A discussion of methods of vocational guidance and selection, including mental tests.

Hollingworth, Leta S. *Special Talents and Defects.* New York: Macmillan Co., 1925.
A description of talents and defects with methods of diagnosing them.

Hull, Clark L. *Aptitude Testing.* Yonkers-on-Hudson, New York: World Book Co., 1928.
A discussion of the theory of aptitude testing and proposed methods.

Hunt, Thelma. *Measurement in Psychology.* New York: Prentice-Hall, Inc., 1936.
A description of tests for the measurement of various physical and mental traits.

Intelligence Tests and Their Use. Twenty-first Yearbook of the National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Co., 1922.
Part I contains a series of general articles giving descriptive and critical discussion of mental tests and of statistical methods. Part II contains articles on the use of tests in various types of schools.

Kuhlmann, F. *A Handbook of Mental Tests.* Baltimore: Warwick & York, Inc., 1932.
A manual for the administration of the Kuhlmann revision of the Binet-Simon scale.

Peterson, Joseph. *Early Conceptions and Tests of Intelligence.* Yonkers-on-Hudson, New York: World Book Co., 1925.
Gives a very full account of Binet's early experiments with tests, of his three scales, and his interpretation of the nature of intelligence tests.

Pintner, Rudolf. *Intelligence Testing.* New York: Henry Holt & Co., 1931 (revised).
A comprehensive treatment of the nature, development, and application of intelligence tests.

Review of Educational Research. *Tests of Personality and Character*, Vol. II, June, 1932; *Tests of Intelligence and Aptitude*, Vol. II, October, 1932; *Psychological Tests*, Vol. V, June, 1935.

Spearman, C. *The Abilities of Man.* New York: Macmillan Co., 1927.
A discussion of the theories of the constitution of abilities and a detailed description of the author's own theory.

Spearman, C. *The Nature of "Intelligence" and the Principles of Cognition.* London: Macmillan & Co., Ltd., 1927.
An account of the psychological principles which underlie the author's theory of abilities.

Symonds, Percival M. *Diagnosing Personality and Conduct.* New York: Century Co., 1931.
A classified description of tests of personality and a discussion of their validity and reliability.

Terman, Lewis M. *The Intelligence of School Children.* Boston: Houghton Mifflin Co., 1919.
Detailed account of the results of the application of the intelligence test to school children.

Terman, Lewis M., and Merrill, Maud A. *Measuring Intelligence.* Boston: Houghton Mifflin Co., 1937.
A manual for the administration of the second Stanford Revision of the Binet-Simon scale.

Terman, Lewis M., and Others. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore: Warwick & York, Inc., 1917.
An account of the standardization of the Stanford Revision.

Thorndike, Edward L., and Others. *The Measurement of Intelligence.* New York: Teachers College, Columbia University, 1927.
A discussion of the theory of measurement of intelligence and a description of the author's investigations in this field.

Thurstone, L. L. *Primary Mental Abilities.* Monographs of the Psychometric Society, No. 1. Chicago: University of Chicago Press, 1938.
A description of the procedure by which the author analyzed intellectual ability.

Whipple, Guy Montrose. *Manual of Mental and Physical Tests.* Baltimore: Warwick & York, 1915 (second edition).
Detailed manual for the administration of fifty-one individual tests, classified according to mental processes, with summaries of results.

Woodrow, Herbert. *Brightness and Dullness in Children.* Philadelphia: J. B. Lippincott Co., 1919.
An account of the characteristics of bright and dull children with emphasis on the findings of mental tests.

Yerkes, Robert M., (Editor). *Psychological Examining in the United States Army.* Washington: National Academy of Sciences, Vol. XV, 1921.
A detailed, technical account of the derivation and application of mental tests in the army.

Yerkes, Robert M., and Foster, Josephine Curtis. *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick & York, 1923 (revised).
A revision and extension of the earlier point scale, published in 1915. The present revision includes scales for young children and adolescents.

Yoakum, Clarence S., and Yerkes, Robert M. *Army Mental Tests.* New York: Henry Holt & Co., 1920.
A brief account of the army tests and guide to their administration.

Chapter II

EARLY EXPERIMENTATION WITH TESTS

1. Early studies of individual differences

THE first clear case on record of the scientific recognition of individual differences in mental abilities occurred in connection with the work of the Greenwich Astronomical Observatory, in England. In 1795, one of the observers was found to differ from his colleagues in his estimate of the time of transit of a star. The method which was pursued was to watch the star through a telescope and to note the time at which the image of the star crossed a line in the field of view. This required that the observer should watch the star until it approached the line and then look at a clock and estimate the exact time at which the star crossed the line. Because this particular observer differed from his colleagues, he was discharged from the staff.

It was later discovered, however, that it is incorrect to assume that some observers are right and others wrong. It was found that there is an error of observation in the case of all observers, and that the amount of this error differs with different individuals. This difference in the amount of error was called the *personal equation*. This term, which first referred to individual differences in the reaction time of observation, came later to be applied to differences in all sorts of mental attitudes and was adopted into general use.

In 1822 astronomers came to recognize the difference in the reaction time of observers and to make allowance for it, but the systematic study of individual differences was not made until more than fifty years later. When psychological laboratories were first founded, the interest, as was remarked

in the first chapter, was mainly in general laws or general principles of human behavior. An illustration of one of the earlier generalizations is the Weber-Fechner Law. This law concerns the relationship between the intensity of a stimulus and the amount of increase in the stimulus which is necessary in order that the person shall detect a difference. In general, the principle is that if a stimulus is very intense, it is necessary to add a large increment in order that a difference shall be perceived, whereas if the stimulus is very faint, a small difference will be perceived. This principle holds for all the different senses, but the proportionate amount which must be added so as to be perceived differs somewhat among the various senses. For example, if one is looking at two lights which differ from one another by a slight amount, it is necessary that the one shall be one per cent more intense than the other in order that the difference may be detected; if the lights are seen in succession, on the other hand, one must be ten per cent more intense than the other. In the case of weights, one must ordinarily be about five per cent heavier than the other, in order that the difference may be distinguished.

It was with such general laws as this, which deal with facts of behavior common to all persons, that the earlier scientific studies in psychology were concerned. It became apparent before long, however, that the differences in the behavior of different persons were of such importance that they could not be neglected. It had been common to designate the differences among the individuals of a group of observers by the term *probable error*. This term represents approximately the average amount by which the reactions of the various individuals differ from the typical reaction of the group. It became evident, in the course of experimentation, that these divergencies of individuals from the other individuals of the group are not due mainly

to error, but constitute real differences in their mental capacity or modes of behavior.

It is interesting that one of the earliest forms of behavior, in which these individual differences were clearly recognized, was the same which occasioned the recognition of the personal equation in the astronomical observatory. This form of behavior is reaction. Cattell, working in the laboratory of Wundt, discovered there were characteristic differences in the reaction time of different persons. This undoubtedly called his attention to the necessity of studying individual differences, and stimulated his later experimentation with mental tests, which proved the starting-point of the development of these tests in the United States.

Another source of interest in individual differences is to be found in the studies of heredity among the English school of scientists. A group of men, including Charles Darwin, Wallace, Huxley, and Spencer, were students of evolution, and of the inheritance of physical characteristics as one phase of evolution. Francis Galton, who was a cousin of Charles Darwin, became interested in the extension of this study to the inheritance of mental characteristics. He believed that temperamental traits and differences in intellectual capacity are inherited in the same way as are physical traits. Galton made a number of scientific investigations to produce evidence on mental inheritance.

In order that the inheritance of differences might be studied, it became necessary to discover means of measuring these differences. One of the methods which Galton used was the questionnaire. He investigated differences in the vividness or accuracy of mental imagery by asking a large number of persons to report what they could remember of the appearance of their breakfast table. In addition to this method, he developed certain instruments for the study of differences of sensation. One of these was the so-called

Galton whistle, which is designed to measure the highest tone which it is possible for a person to hear.

Cattell was for a time associated with Galton, and this association strengthened the interest in individual differences which had been developed from his earlier study. Upon taking up his work as a teacher of psychology in the United States, Cattell proposed a program of tests. This program was published in the British journal, *Mind*, in an article which was published[1] in 1890. Galton contributed a number of comments at the end of the article, and this gives us direct evidence of the connection between the study of inheritance in England and the mental testing movement in the United States.

The purpose of the program of tests which was set forth in Cattell's article was as follows: first, to determine the constancy of mental processes, or the degree in which they vary from time to time in the same individual; second, to determine the degree of interdependence between the various mental processes; and, third, to determine the amount of their variation under different circumstances. While the tests constitute means of measuring the differences between the behavior of different persons, it will be seen that a considerable share of the interest in them was still concerned with the analysis of the mental life of the individual considered alone.

The character of these early tests may be gathered from the list of ten which were recommended for most extended use. They were as follows:

1. Measurement of the strength of grip by the dynamometer. This is an instrument containing a strong spring which is compressed by the grasping movement of the hand. The amount of the pressure which is exerted is recorded upon a dial.

[1] J. McK. Cattell, "Mental Tests and Measurements," *Mind*, XV (1890), 373–80.

2. Measurement of the rate of movement. This consisted in the measurement of the quickest possible time in which a person could move the hand through fifty centimeters.
3. The measurement of the smallest distance between two points placed on the skin which can be distinguished as two by the individual. This measurement was made by an instrument called the esthesiometer.
4. The measurement of the amount of pressure necessary to cause pain. The pressure was exerted upon the forehead by a strip of hard rubber.
5. The measurement of the smallest amount of difference in weight which can be discriminated. The measurement was made by requiring the subject to lift two weights in succession.
6. The measurement of the quickness with which a person can react to a sound.
7. The measurement of the quickness with which a person can name ten specimens of four different colors arranged in miscellaneous order.
8. The accuracy with which a person can bisect a fifty-centimeter line.
9. The accuracy with which the individual can reproduce an interval of ten seconds. The subject responded by giving a signal to mark the termination of an interval of time which he judged equal to one which was marked off for him by two previous signals.
10. Immediate rote memory. The number of consonants spoken to an individual which he can repeat immediately afterward in series.

It will be seen that these tests measure acuity of sensation, rapidity of movement, simple judgment, and simple memory. We shall find it instructive to keep in mind the character of these early tests and to compare them with those which were employed during the succeeding decade, and with those which were developed during the period since 1900.

2. Early American experiments with tests

In the few years following 1890, Cattell began to apply

mental tests to students in Columbia University. These tests were continued in systematic fashion until the end of the decade, and were reported upon in a monograph by Clark Wissler, in 1901. We shall return shortly to some of the results which were reported by Wissler.

The type of interest in mental tests during this period is illustrated by the activities of the American Psychological Association. At the instance of Cattell, the Association appointed a committee, in 1895, for the purpose of formulating mental tests, and of developing a program for their use. The committee consisted of some of the most prominent psychologists of the country. In 1896, the committee presented to the Association a long list of tests. It was recommended that they be given to college students. They were chosen for the purpose of measuring intellectual growth and individual differences. It was believed that they could be given in one hour. They were, of course, to be given individually.

It is not necessary to reproduce in detail this list of tests. They were simply elaborations of the list which had earlier been proposed by Cattell in his article in *Mind*. The general character of the mental capacities which were the subject of the test was the same. There were only one or two tests of a more elaborate type, which were designed to measure the ability to react to a complex situation or to carry on the processes which we ordinarily designate as judgment, thinking, association, or the higher forms of memory for complex materials.

The committee of the Psychological Association was continued for several years, but, so far as the record goes, did little more as a committee than to report lists of tests of this character. The most elaborate use of the tests was made at Columbia University. A number of other psychologists gave the tests to college students or to other persons.

For example, during the World's Columbian Exposition at Chicago, in 1893, Jastrow had a booth at which he gave a series of tests to persons who offered themselves as subjects. Jastrow had previously reported a list of tests which had been given college students as early as 1891. Most of these early reports, however, contained no statements, or very meager statements, concerning the results of the tests.

A few sporadic tests were given school children during this early period. The character of the tests, their aims, and their results may be gathered from a few typical illustrations. A memory test, consisting of the measurement of the memory span for digits, was given to a group of school children by T. L. Bolton, in 1891. The scores which the children made in these tests were compared in a very crude way with the estimates of their teachers concerning their general mental ability. The experiment thus anticipated, after a fashion, our modern method of examining and trying out tests.

The results of this comparison are shown in Table I.

TABLE I. THE RELATIONSHIP BETWEEN TEACHERS' ESTIMATES OF MENTAL ABILITY AND MEMORY OF DIGITS [1]

Classification by Teachers' Estimates	Classification by the Memory Test: A	B	C
A	32.6%	51.0%	16.3%
B	21.4%	58.2%	20.4%
C	24.1%	49.4%	26.5%

On the top line of the table, opposite the letter *A*, are

[1] Thaddeus L. Bolton, "The Growth of Memory in School Children," *American Journal of Psychology*, IV (April, 1892), 362–80.

indicated the children who were rated the highest by the teacher. In the middle row are those who were rated as average, and in the bottom row those who were rated as poor. In the vertical columns are represented the percentages of the children of these various groups, based on teachers' estimates, who fell into the three groups on the basis of the memory tests. For example, to begin with the upper left-hand figure, we see that 32.6 per cent of the children who are rated in the top third by the teachers fell also into the top third in the memory test, 51 per cent of these children fell into the middle third in the memory test, and 16.3 per cent in the lower third. Evidently, then, there was some relationship between the teachers' estimates and the tests, although that relationship was not at all close. If we examine the second and third rows, we find that the correspondence between the memory-test scores and the teachers' judgments of the abilities of the children in the two lower groups is very slight.

This method of measuring the relationship between a test and some other measure of a child's ability is, of course, a rough and crude method. It will be found to be in sharp contrast with the more refined methods which were later put into operation. So far as the comparison may be relied upon, however, it indicates that the effort to discover a means of testing the child which would agree with the estimate made of him by his teachers was almost a total failure.

A somewhat similar comparison between the standing of children in certain mental tests and the estimate of their ability by their teachers was reported by Gilbert, in 1894.[1] This comparison of the tests with teachers' estimates was rather incidental to the main purpose of Gilbert's study.

[1] J. A. Gilbert, "Researches on the Mental and Physical Development of School Children," *Studies from the Yale Psychological Laboratory*, Vol. II, pp. 40–100. 1894.

His purpose was to measure the growth of children in a variety of mental capacities by giving tests to one thousand children of different ages. For our present purpose we may confine ourselves to his comparison of the standing of three groups of children. He asked the teachers to divide the children into three groups which should be called bright, average, and dull. He gives us the average scores made by these three groups in each of the tests. The tests include chiefly measurements of reaction time, simple memory, and various types of sensory discrimination. We may take, as a single example, the comparison of the average reaction time of the three groups. They are as follows:

Bright	Average	Dull
20.7	21.2	22.4

If we may take the averages as a reliable indication of the differences between these groups, we may say that there is a slight difference in favor of the bright children. The smaller reaction time, of course, indicates the higher score. We must interpret this difference, however, in the light of the variations which we find in each group. The average mean variation which is reported by Gilbert is 3.6. This is over twice the difference between the average of the bright group and the dull group. It indicates, therefore, that the difference between the groups is so slight in comparison to the differences between the individuals in each group that it would be at least of no diagnostic value and is of little significance of any sort. We must, of course, raise the question whether the teachers' estimates gave a reliable classification.

Certain pertinent questions regarding the technique of making these judgments are not answered in the report. For example, did the teachers divide the classes into equal groups? Did they allow, in making their judgment, for the

differences in age of the pupils they compared? Did they have an adequate idea of what was meant by brightness — did they distinguish between natural ability to do school work and the actual attainment of the children, which we know do not always agree with native ability. Later experiments with tests have indicated that it is very necessary to take such matters as these into account, and in the most careful manner, if the statistics which result from such comparison are to be at all relied upon.

On the face of the returns, the indication is that there is very little relationship between such a mental ability as is represented in the measurement of reaction time and the brightness which is exhibited by children in the school. It may be said at once that while the classification of the children into three groups was probably not made as accurately as could be desired, the low diagnostic value of such a simple test as that of reaction time is confirmed by other later experiments such as the one reported by Seashore, in 1899.[1] Seashore gave a number of tests to a group of school children, among them tests of sensory keenness. These consisted of keenness of hearing, discrimination of pitch, and time memory. He reported that there was no correlation between the standing of the children in these sensory tests and their brightness. A later investigator, to whom we shall have occasion to refer at some length, Spearman, recalculated by a more elaborate method the relationship between pitch discrimination and brightness as reported by Seashore and reports a correlation of .20. This correlation, even though it indicates some relationship, is so small as to be of no practical importance. That is, a test which has such a low correlation as this could not be used as a means of diagnosis of intellectual capacity.

We may take, as our final illustration of the early applica-

[1] C. E. Seashore, *Some Psychological Statistics.* University of Iowa Studies. 1899.

tion of tests to American school children, a study of the relationship between motor ability and marks which was reported by Bagley, in 1901.[1] Bagley applied tests to measure the following characteristics of movement: strength, rapidity of voluntary movement, accuracy of voluntary movement, steadiness of motor control, amount and character of involuntary movement, and reaction time. The results from all of these tests taken together were combined to form what was called a motor index. The marks which the children made in their school subjects were then averaged and taken to represent class standing.

The comparison between the motor index and the class standing was made in the following manner. The motor indices were first divided into five equal-sized groups, according to rank. The average of each of these five groups was then calculated. This gave a descending series of averages. The average class standing of the children whose scores appeared in each of these five groups was then found. The two series of averages are given in the following table:

A Comparison of the General Motor Index and Class Standing

Motor index	961.8	938.3	924.3	909.0	881.9
Class standing	77.8	80.0	83.6	83.8	84.7

The scores were then reclassified in the reverse manner; that is, the school marks were first divided into five groups according to rank, and the averages of the groups were calculated. The average motor index of the children represented in these groups, classified on a basis of class standing, was then found. The two series are given below.

Class standing	92.7	87.5	83.0	74.0	67.9
Motor index	917.2	907.1	922.8	931.0	931.0

[1] W. C. Bagley, "On the Correlation of Mental and Motor Ability in School Children," *American Journal of Psychology*, XII (January, 1901), 193–205.

It will be seen from these comparisons that the children who stood high, on the average, in motor ability stood comparatively low in class standing, whereas those who stood low in motor ability stood higher in class standing. There is apparently what we now call a negative correlation between motor index and class standing. Later studies indicate that such an opposition between motor ability and school marks does not exist. This study, while it was a carefully conducted experiment in comparison with other studies of the time, shows the necessity of observing certain precautions in statistical procedure in making tests and in making comparisons from their results. These precautions have been discovered by experience, and can be said now to constitute a body of technique which is characteristic of the modern procedure in testing.

The probable explanation of the negative result of the comparison which was made in the study by Bagley is that the children who were compared with reference to their class standing and to their motor ability differed in age, but were not representative of all the children of their respective ages. The older children of a class, for example, are known to be lower in general academic ability than the younger children. This is because the older children are the ones who are retarded, due to their dullness, and the younger ones those who are accelerated because of their unusual brightness. The older children, because of their mere age, on the other hand, are superior to the younger children in their motor development. This superiority of the older children in motor development, coupled with their lack of superiority or actual inferiority in academic ability, produces the appearance of inverse correlation or of opposition between motor abilities in general and class standing. While later studies have indicated that the relationship between motor ability and general academic ability is slight, they do not confirm this finding of a negative correlation.

This study, then, illustrates two points. First, if any relationship between motor ability and general academic ability exists, it is very slight, and motor ability therefore is not a suitable subject of testing if we wish to measure general intellectual capacity. Second, it is necessary to adopt the most careful technique in the administration and interpretation of the results of tests. We shall have to consider the demands of technique in considerable detail in the course of our later discussion.

We may now return to the Columbia tests and close our account of the experimentation in the United States during this early period by a summary of the results of the experiments reported by Wissler, in his monograph in 1901.[1] Because of the historical importance of these Columbia tests, and in order to indicate in somewhat more detail the character of the early tests, we may reproduce the list of the Columbia tests in full. The list of traits or capacities which were measured in the Columbia tests is as follows:

Length and breadth of head.
Strength of hand.
Fatigue as measured by an instrument called the dynamometer.
Acuity of vision.
Color vision.
Acuity of hearing.
Pitch discrimination.
Weight discrimination.
Discrimination of two points on the skin by the esthesiometer.
Pain sensation.
Perception of size.
Color preference.
Reaction time.
The rate of the perception and reaction as measured by the rapidity of crossing out *a*'s in a text.

[1] Clark Wissler, *The Correlation of Mental and Physical Tests.* Psychological Review Monograph Supplements, Vol. III, No. 6. Lancaster, Pennsylvania: Macmillan Co., 1901.

The rapidity of naming colors.

Rate of movement as measured by dotting in one centimeter squares with a pencil.

Accuracy of movement as measured by striking dots with a pencil.

Perception of time as measured by the ability to follow rhythm one second after the sound has ceased.

Association as measured by free associations to nine words.

Imagery as measured by the imagery test of Galton.

Memory as measured by four simple memory tests.

The memory tests involved the immediate repetition of numbers which were seen, the immediate repetition of numbers heard, the repetition of a passage, and the ability to remember the length of a line which had been seen in the early part of the test period.

It will be seen that the character of the tests is similar to the character of the early tests which were listed by Cattell in his article in *Mind*. They are somewhat more elaborate in that a larger number of tests are used, but they are nearly all tests either of the accuracy of sense discrimination, or the strength or rapidity of movement. The last three tests, those of association, imagery, and memory, are somewhat more complex in nature than the others. They give a suggestion of the type of test which has since predominated. They were, however, very incompletely developed, as compared with the tests which are in present use.

The results of the tests which are of most interest to us concern the correlation between the various tests themselves and the correlation between the standing in the tests and college marks. The degree of correlation is expressed in a much more accurate fashion than in the earlier studies to which reference has already been made. The meaning of correlation will be explained more fully in the next chapter, but for the present we may merely say that it represents the closeness of correspondence between two traits. Correla-

tion is represented as a coefficient. This coefficient may range from -1 to $+1$. Perfect agreement is represented by $+1$. If there is no ascertainable relationship between the scores in two tests, except what would be present by mere chance, the coefficient of correlation is zero. If the relationship is reversed, so that high standing in one test corresponds with low standing in the other, the coefficient becomes negative, and if this relationship is the extreme opposite of complete correspondence the coefficient becomes -1.

With this brief explanation, we can understand the significance of the coefficients which are reported. A few may be selected as typical examples. We may first mention a few correlations between the various tests themselves.

TABLE II. CORRELATION BETWEEN CERTAIN OF THE COLUMBIA TESTS

Reaction time and naming colors	.15
Reaction time and association	.08
Marking *a*'s and naming colors	.21
Speed of movement and naming colors	.19
Speed of movement and reaction time	.14
Reaction time and marking *a*'s (approximately)	0

We see from this list that the relationship between the scores in the different tests is very low. The highest of these correlations, .21, represents only a slight degree of relationship. The degree of correspondence which is represented by such a coefficient is so slight that one could not use the score in one test in any practical way to predict what the score of the individual in the second test would be. On the face of it, therefore, the abilities which are represented by the scores in these tests have very little relationship to one another.

We may contrast the results of the mental tests with the two characteristics, height and weight. The correlation

between these was .66. This means, for example, that if we knew which quarter of the entire group a person belonged in with reference to height, we could more often than not predict which quarter in reference to weight he would also belong in. This, of course, is not an extremely close correspondence, but it is close enough to make it possible to use the scores in one test to predict with enough accuracy for certain practical purposes what the score will be in another. The correlations between the various tests, however, are not sufficiently accurate and related to make this possible.

The next comparison that we may make is between the standing in the tests and in the college classes. The standing in each test was compared with the average of the marks in all of the courses taken by the individual student. This average class standing was found to be correlated with the standing in the test, as follows:

TABLE III. CORRELATION BETWEEN AVERAGE CLASS STANDING AND A NUMBER OF MENTAL TESTS

Class standing and reaction time	−.02
Class standing and marking *a*'s	−.09
Class standing and association time	.08
Class standing and naming colors	.02
Class standing and logical memory	.19
Class standing and auditory memory	.16

It appears that the tests were not more closely related to class standing than to each other. The highest correlation is between class standing and logical memory, and this is negligible so far as the use of the tests for diagnosis is concerned. In contrast to these low correlations is the correlation between the standing in the various subjects themselves and between average class standing and gymnasium work. These correlations are given in the next table.

Table IV. Correlation between the Standing in the Various College Subjects

Latin and Mathematics	.58
Rhetoric and Mathematics	.51
Rhetoric and Latin	.55
Rhetoric and French	.30
Rhetoric and German	.61
Mathematics and German	.52
Latin and French	.60
Latin and German	.61
Latin and Greek	.75
Average class standing and gymnasium grades	.53

It is evident that the low correlation which was found between the tests themselves and between the tests and class standing is not to be explained by a deficiency in the technique of finding correlation. The high correlations between the standing in the college subjects preclude this explanation. The reason for the low correlation must be found in the nature of the tests themselves.

There are two possible explanations of the low correlations. They may be due either to the nature of the mental processes which are being tested, or to the faults in the technique of the organization or administration of the tests. Before we can go fully into the discussion either of the content of the tests, or of the technique of their organization and administration, it is necessary to discuss these matters more fully than is possible at this point. We must content ourselves for the present, therefore, with merely hinting at the possible explanation of the negative results of these tests. The fuller contrast between these earlier tests and the subsequent ones will be more thoroughly appreciated after the later stages in the development of tests have been described.

For the present, it is sufficient to say that the low correlations which are here reported were probably due, in part, to the fact that the mental processes which were tested were

chiefly the sensory and motor processes. Nearly all of the experimental work with tests has demonstrated the fact that these mental abilities are not closely related to one another, or to the complex activities which comprise achievement in the school, or to achievement in vocational activity outside the school. The low correlations may also be due in part to the fact that the tests must have been given somewhat hastily, since the entire series was completed within an hour. It is possible, therefore, that an individual's score in any particular test was not a stable measure of his capacity. If the test had been given a second time, his standing might have been altered seriously. Since the consistency of the tests was not measured, we cannot say how far this supposition is correct in this particular case. Later experience, however, indicates that the scores in individual tests are often not very constant. If this is true, it explains in part the low correlations between tests, and between the tests and college marks.

3. Early European experiments with tests

We have already seen how the interest in mental tests as measures of individual differences was the outgrowth of the work of psychologists in the experimental laboratory. A number of European psychologists were carrying on experiments with tests during the decade from 1890 to 1900, which paralleled, in the main, the experiments of American psychologists. Among the most prominent of these Europeans, both because of the amount of work which he was doing at the time and because of the importance of the later outcome of his experiments, was the French psychologist, Alfred Binet.

It is instructive to notice that Binet's earlier work was apparently as lacking in immediate productiveness as was that of the American psychologists. He was using very

much the same kind of tests, and interpreted their results with very much the same sort of rough methods. There are to be found in his work, however, germs of the characteristics which were responsible for the success of his later experiments. These appear in the practical character of his aims and interests, and in some measure in the character of the mental processes which he was trying to measure and which were represented in the tests of his later scale. We may glance briefly at two of the articles in which he reported his early work.

In 1895, Binet[1] proposed a list of tests, much as did the American psychologists of the same period. He did not report the result of the applications of the tests, nor did he report in detail methods by which the tests could be scored. His publication, therefore, was lacking in immediate productiveness. It is interesting, however, as indicating that Binet was experimenting with tests somewhat different in character from the tests of sensation and of movement which characterized the American work. We may illustrate by a few examples from the list which he presented. Binet first suggested four tests of memory. Two of these were later used in his scale. One was the memory of geometrical designs, a second tested the memory for a short paragraph, and a third tested immediate memory for numbers. A second test was designed to measure the character of the individual's mental images. Another series of tests was designed to measure attention, either the uniformity of attention or the number of the objects or ideas which could be kept in mind at one time. Another group of tests were to measure what Binet described as comprehension. Other groups were designed to measure suggestibility, æsthetic feeling, and moral sentiments.

[1] A. Binet and V. Henri, "La psychologie individuelle," *Année psychol.*, t. 2 (1895), 411–65.

It will be seen that Binet's proposals, in contrast to those of some of his fellow psychologists, were very vague, and that the exact means by which the abilities which he described were to be tested had not been worked out by him. He had not as yet developed a technique to measure some of the functions which he listed. The success of the testing movement, however, has been due in part to the efforts to measure some of these more complex mental processes, in contrast to the simpler ones studied in the early American work.

Binet himself experimented with a number of tests of the simpler type, which were devised and given by him for the purpose of measuring attention and adaptation.[1] The practical cast of Binet's mind was indicated by the fact that he gave these tests to two groups of children, six being the poorest from a class of thirty-two, and five being the best. He selected these two groups in order that he might determine which tests served to differentiate the bright from the dull pupils.

The first test measured tactual discrimination by means of the esthesiometer. This test was given in three forms. In the first form the bright children excelled the dull ones. In the last form, however, which was somewhat more accurate and in which there had been opportunity for some practice, the difference between the two groups was very slight. The second test was a measure of reaction time. Binet's results agreed with those of Gilbert in that there was little difference between the scores of the two groups of children. The next test consisted of counting small points placed close together. Here again the difference between the two groups was small or non-existent. In the next test the child listened to the count of the beats of an instru-

[1] A. Binet, "Attention et adaptation," *Année psychol.*, t. 6 (1899), 248–404.

ment called the "beater." He was required to tell whether or not, at a predetermined given time, the rate of the beater was changed. In this test the dull children exceeded the bright ones. This result was explained by Binet as due to the fact that the children knew when to pay attention. He believed if the warning had not been given, the bright children would have exceeded the dull ones.

In the remaining tests the bright children, in general, excelled. The first one consisted in counting the beats of a metronome. Here the bright children made higher scores than the dull ones, though the dull children improved more than the bright children. The next test consisted in copying various sorts of printed material, and the measure was the amount which could be copied at one act of observation. The bright children in each case excelled the dull children in this test. The bright children also made fewer errors than the dull ones in reproduction of letters or numbers from memory. In the reproduction of a design seen for a brief space of time, the bright children gave better reproduction than the dull ones. In the test of crossing *a*'s, the dull children made more errors than the bright ones. A similar result came from the test in simultaneous adding. The speed of the two groups was approximately the same, but the number of errors was greater in the case of the dull children. In speed of reading and copying sentences, the dull children were equal to the bright ones.

This experiment would, of course, not at all meet our present statistical standards for an experiment in tests. Its results can be regarded as merely suggestive. If we take the results of the comparison of the two groups of children as a whole, we see that those tests in which the mental processes were more complex did, in general, differentiate the two groups better than those in which the mental processes were simpler. This undoubtedly impressed Binet and led him to

the selection of the more complex type of tests for his later scales.

Little work of any note, aside from that of Binet's, was done in Europe during this early period. Some experimentation with tests was stimulated by Kraepelin, who was interested primarily in means of diagnosing insanity. A. Oehrn,[1] in 1895, published a report of a few tests and gave data from which their correlation could be calculated. These tests dealt with perception, which was measured by counting letters, crossing letters, and proof reading; with memory; with simple association processes, such as are required in adding; and with several motor functions, such as writing from dictation and reading simple material. Krueger and Spearman calculated the correlation between the scores in these tests, and found them to range from .44 to .69, aside from the cases in which no correlation was found. This experiment is interesting merely because it was representative of the earlier work, and proved somewhat more successful than some of the other experiments which were made at this time.

A marked exception to the unsuccessful attempts of the earlier tests was the invention of the so-called "completion test" by Ebbinghaus, in 1897.[2] Ebbinghaus aimed primarily to find a measure of intellectual fatigue, and he therefore selected a mental process which he thought would represent the higher intellectual activities. His analysis led him to believe that the combining activity of the mind was the highest. In order to measure this activity, he devised a test in which the subject was shown a text with certain words left out. He believed that if the individual had the capac-

[1] A. Oehrn, "Experimentelle Studien zur individuelle Psychologie," *Psychol. Arbeiten*, B. 1 (1895), 92–151.

[2] H. Ebbinghaus, "Ueber eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern," *Zeitsch. f. Psychol.*, B. 13 (1897), 401–59.

ity in general to put together the items of his experience in such a way as to see their relationship, this capacity would be measured by his score in filling in these blanks. While his device did not prove very successful as a measure of general intellectual fatigue, it did prove valuable as a measure of general intellectual capacity. It was tried out with children of different levels of school achievement in Germany, and has later been very extensively used by European and American psychologists.

The work of Francis Galton in the investigation of individual differences, for the purpose of tracing the inheritance of mental characteristics, has already been mentioned. While this work did not have much direct influence upon testing in the schools, it did have a large amount of indirect influence. It is interesting to notice in this connection that Galton's successor in the Eugenics Laboratory in London, Karl Pearson, formulated a method of calculating correlation which is now in wide use in the field of mental tests.

The interest in individual psychology was also promoted by the work of William Stern.[1] While Stern did not, in this early period, develop mental tests, he made a study of the intellectual characteristics of prominent men — men who were known for unusually high intellectual achievement in some line of endeavor. He attempted to analyze the special capacity which was possessed by these men by the biographical and questionnaire methods. Stern later became interested in the application of tests of intellectual ability.

We have now reviewed briefly the typical tests which are characteristic of the early period, which closed about 1900. We may summarize briefly the outstanding characteristics of these early tests.

[1] W. Stern, *Ueber Psychologie der individuellen Differenzen.* 1900.

4. Summary of the early period

The interest in tests during this period was largely theoretical. It was the outgrowth of the work of the psychologist in the psychological laboratory. It was related to the general interest in individual differences, and this again was related to the inheritance of mental traits. There were, to be sure, some experiments in the application of tests to school children, and these experiments foreshadowed, in a measure, the practical applications which have been made of tests during the past two decades.

Most of the tests of the early period, as has already been pointed out, were single tests. They were not organized into scales. If a number of tests were given at the same time, the scores which were made in these tests were not combined. This again constitutes a marked contrast between the earlier tests and most of those which are now in use.

In the next place, the early tests were not standardized in the sense which we use the word standardized in the present time. No careful method was used to determine whether or not the tests were reliable. Reliability may depend upon the way in which the test is given, or the way in which the response by the individual is scored, or upon the general conditions under which the test is taken. We now have careful methods of determining whether or not a test is thoroughly standardized. Furthermore, certain practices have been determined upon as constituting good standardization. In this respect again, then, the early tests are in contrast with those which were later developed.

In addition to the fact that there was no careful procedure by which the reliability of a test was determined, there was no well-recognized method for accurately determining the significance of a test by comparing the scores made in it with other measures of achievement. Some comparisons, to be

sure, were made, such, for example, as those of Bolton, of Gilbert, and of Binet, but these comparisons were made on a small scale in many cases, and without the elaborate statistical technique which is now customary.

So far as the content of the early tests was concerned, it dealt mostly with the sensory and the motor processes. In some cases simple tests of memory were used, and in a few cases, such as some of the tests of Binet, and the Ebbinghaus test, the higher mental processes were included in the measurement. These, however, are exceptions to the general rule.

As there was no highly organized method of comparing the results of tests with other methods, so the method of systematic comparison of results was not used as a means of selecting the tests. The selection of tests, on the contrary, was based largely upon a preliminary analysis of the mental process which it was desired to measure. This preliminary analysis is not to be criticized in itself. It has, however, been shown to be inadequate as a sole method or criterion for selection. It is probably true that the further advancement of tests will depend in a measure upon a more acute analysis than we have been able to make at this time of the mental capacities. However, this analysis must be backed up by a careful statistical examination of the result.

Finally, the results of the early tests, so far as we may judge by the comparison with other measures of achievement, were for the most part negative. This negative outcome is illustrated, for example, in Wissler's report of the Columbia tests. The consequence of this negative outcome is that tests for a time fell into a distinct disfavor on the part of professional psychologists. There was some experimentation with tests during the succeeding few years, but this was sporadic and did not elicit the interest of psychologists as a body. The type of prevailing interest on the part of professional psychologists is well illustrated by the ap-

pointment of a new committee on tests, in 1906.[1] This new committee, in contrast to the earlier one, took as its purpose not the organization of an elaborate series of tests but the minute standardization of a few tests of simple sensory and motor processes. The work of this committee made a valuable contribution to the measurement of these simple processes, but did not contribute towards the development in the direction of testing the more complex or the higher mental activities.

In the next chapter we shall discuss the further development of single tests, particularly as it is related to the correlation technique.

[1] The work of the committee was reported in the monographs given in the list of references which follows:

REFERENCES

Report of the Committee of the American Psychological Association on the Standardizing of Procedure in Experimental Tests. Psychological Monographs, Vol. XIII, No. 1. Lancaster, Pa.: Review Publishing Co., 1910.

This report contains the following parts:
(*a*) Methods for the Determination of the Intensity of Sound, W. B. Pillsbury.
(*b*) The Measurement of Pitch Discrimination, C. E. Seashore.
(*c*) The Determination of Mental Imagery, James R. Angell.

Woodworth, R. S., and Wells, Frederic Lyman. *Association Tests.* Psychological Monographs, Vol. XIII, No. 5. Lancaster, Pa.: Psychological Review Co., 1911.

Yerkes, R. M., and Watson, John B. *Methods of Studying Vision in Animals.* Behavior Monographs, Vol. I, No. 2. New York: Henry Holt & Co., 1911.

Chapter III

THE APPLICATION OF THE CORRELATION METHOD

Methods of investigating correlation had been used in the early period of the development of tests. These methods, as has already been pointed out, were crude, and they had the defect that it was not possible to establish the degree of correlation by means of a single numerical quantity. As we have seen, there was one investigation in which the coefficient of correlation was calculated, namely the study by Wissler. In this case, the correlation was calculated only after the tests had been given. It was not used in the development of the tests themselves.

We may pause for a moment to explain in brief the general meaning of correlation. It expresses the degree of correspondence between two traits, such, for example, as height and weight, or musical ability and artistic ability. It is only possible to measure such relationships by a comparison of the amount of ability possessed by the various individuals of a group. In other words, correlation is entirely a comparative affair.

We may take as a simple example the two characteristics of height and weight. If all the individuals of a group were compared in both these respects, it would be found that on the whole the taller persons are also the heavier. If, now, this correspondence were complete there would be a perfect correlation between height and weight. One way of expressing this correspondence is to say that the tallest person of the group is also the heaviest, and the shortest person the lightest. This expresses the relationship in terms of rank-order. We may, however, make a somewhat more exact

comparison by calculating the average height and weight of the members of the group, and then representing the height and weight of each individual in terms of the degree of variation above or below the average. If, now, an individual varies from the average in height by an amount which corresponds exactly to his variation from the average in weight, and this is true of all the individuals in the group, we say that the correlation is perfect.

It might, of course, be *a priori* possible that we should find two traits in which there was not only a lack of positive correspondence, but even an inverse relationship. If the persons who stood at the top of the series in reference to one trait stood at the bottom in reference to the other, and if all the other individuals stood in this relation of opposition, we would then say that there was a complete inverse relationship or negative correlation between the two traits. This represents the other extreme.

Many degrees of correlation may exist between these two extremes. There may be a high degree of positive correlation, but not a perfect one; there may be a high degree of negative correlation; or there may be low degrees of negative or positive correlation. If no determinable relationship exists of either positive or negative character, we must conclude that the relation between the two traits is one of chance.

It is not the place here to enter upon the description or the discussion of the means of calculating the correlation between traits. We may simply refer again to the quantitative expressions which are used to represent the degree of correlation, and which were given on page 48. The foregoing statement may suffice to indicate the nature of correlation for the purpose of interpreting its use in the development of mental tests.

We have set in contrast the earlier and the later tests on

the ground that the former were interpreted by means of rough or crude methods of finding correlation, whereas in the latter the interpretation was based upon more refined methods. Another distinction between the two periods is that while, in the case of the earlier tests, correlation was applied after the tests had been given, and for the purpose of interpreting results, in the later period correlation was also applied while the tests were being developed, in order that it might be determined which tests were good and which were poor. In other words, correlation became a part of the technique in the design and organization of a test.

By means of correlation it is first determined whether a test gives consistent results. For this purpose the scores on the test are correlated with the scores in the same test given a second time. The coefficient which results is called the reliability coefficient. In the second place, the significance or meaning of the test is examined by finding the correlation between the scores on this test and the scores on other tests. This indicates to what extent the various tests measure the same capacities. This is a factor in the interpretation of the meaning of the test scores and in the design of composite scales. Finally, the test is examined by finding the correlation between scores in the test and some outside measure altogether. We shall see how these criteria are applied in greater detail in the course of the discussion.

1. Spearman's criticism of statistical procedure

The advance to the more precise method of standardizing tests and of calculating their results is represented in the writings of Charles Spearman. Spearman, in 1904, published an article[1] entitled "'General Intelligence' Objec-

[1] C. Spearman, "'General Intelligence,' Objectively Determined and Measured," *American Journal of Psychology*, XV (April, 1904), 201–93.

tively Determined and Measured." In this article he reviewed in a critical way the previous tests, outlined the main problems for study, and indicated the technique by which these problems might be attacked. His criticism of the previous work is summed up under four heads. It was defective, he says, first, because the investigators failed to use precise quantitative expressions to represent the degree of correlation between tests, or between tests and other measures. The previous work failed, in the second place, because it did not include a calculation of the probable error of the correlation. In the third place, it did not eliminate certain irrelevant or falsifying factors which might give a misleading correlation which was too high or too low. Finally, in the fourth place, it did not allow for errors in observation. We may review briefly each of these criticisms.

Spearman did not, himself, as we have already seen, invent the mathematical formula for calculating correlation. The formula had already been devised by Karl Pearson, who modified a previous method developed by Bravais. This method had already been used by Wissler, and by Aiken, Thorndike, and Hubbell. Pearson's method was called the *products-moment method.*

While the products-moment method can be applied by one who knows no higher mathematics, it does require rather elaborate calculation. Spearman contributed to the ease of finding correlation by presenting two simpler formulæ. One of them is called *Rank Method,* and is a method of finding correlation by ranking instead of by calculating the variations of individual scores from the average. A shorter or *Footrule Method* is dependent also upon the procedure of ranking. These two simple formulæ of Spearman have been very widely used where the number of cases is rather small and one does not desire a very precise calculation.

Spearman's contribution, however, is not so much in the production of these formulæ as in the emphasis which he placed upon the need of precise methods of calculation. The rougher methods, such as were used by Bolton, Gilbert, and Bagley, can be used to determine whether a marked degree of correlation exists; but to determine whether there is more or less correlation between two pairs of traits, or precisely what the degree of correlation is, it is necessary to express the degree of correlation in a single numerical quantity. It is this which the methods of Pearson and Spearman enable us to do.

The expression of the degree of correlation between traits in a single coefficient also made possible the comparison of the degrees of relationship between the various pairs of a whole group of traits. This led to the study of the interrelationship between mental capacities on a large scale, and also to the attempt to interpret the cause of the interrelationships which are found to exist. Spearman introduced this study, in association with Krueger, in an article which appeared in 1906.[1] He was followed rapidly by several other investigators who made a larger number of tests. We shall have to consider the several studies which grew out of this investigation in a later section.

Spearman emphasized, in the second place, the necessity of calculating the probable error of the coefficient of correlation, or, in more general terms, of determining how reliable a coefficient is. Pearson's formula for calculating the degree of probable error of a coefficient was already in existence, but little use had been made of it. There are two main factors which affect the probable error of the correlation coefficient. In the first place, the number of cases is an important factor.

[1] F. Krueger and C. Spearman, "Die Korrelation zwischen verschiedenen geistigen Fähigkeiten," *Zeitschrift f. Psychol.*, B. 44 (1906), 50–114.

If there are only a few persons in the group for which the correlation is found, chance plays a large factor in the size of the coefficient. To put it in another way, if another group of the same size is tested and the correlation between the two tests found, the probability is very large that it will differ considerably from the correlation of the first group. The error which is caused by using a small number of cases is called the error of sampling.

The second factor which affects the size of the probable error is the size of the correlation coefficient itself. The larger the coefficient is, the smaller the probable error will be. The size of the probable error, then, indicates how much we can rely upon a particular coefficient as being a stable measure of the relationship between the traits which are compared, so far as the merely statistical factors affect it. It is now customary always to calculate the probable error of a coefficient of correlation, and to report it with the coefficient itself. If the correlation coefficient is not at least four or five times as large as the probable error, one should place no reliance upon the coefficient.

The converse of this statement, however, is not always true. We cannot, in every case, rely upon a correlation coefficient which meets these demands of statistical reliability. Because a coefficient is four or five times the probable error does not mean that we can expect to get a coefficient within the range of the probable error in half the cases when we calculate the correlation between the same two traits in the case of another group of persons. There are other variable factors which affect the degree of correlation other than those of a statistical nature. For example, the army test has been applied in a large number of academic institutions, and the correlation has been found between the scores in the army tests and in the marks of college classes. The amount of correlation thus found has varied all the way from about

.3 to .7 or higher. We find, when we come to review the results of the study of the correlation between mental traits, that the variations in the coefficients which are found are confusingly large. These variations make it necessary to be very cautious in the interpretation of the results of correlation, even though we may carefully follow the most reliable statistical method. They make it unsafe to place much reliance upon any single correlation. It is necessary to base our interpretations upon the general trend of correlation coefficients rather than upon any single measure.

The causes of still other variations among correlation coefficients, which cannot be accounted for by an inadequate number of cases, are touched upon in Spearman's remaining criticisms. He says, in the third place, that the calculated coefficient may be affected by other factors than the real relationship between the traits which are measured. It may, for example, be raised or lowered by kinship between the individuals who are tested, by differences or likenesses of the social level of the individuals, and possibly by differences in attitudes or abilities which affect the score but which are not the thing which it is desired to measure, such as zeal, endurance, or manual dexterity. These may either produce an apparent correlation when none exists between the traits, or reduce the correlation coefficient below the true measure.

Let us take age as an illustration of these irrelevant factors, because it is clear that it does frequently affect test scores in such a way that it will produce an error unless it is properly accounted for, and because it is a factor which is often overlooked. Table V[1] illustrates the way in which age may increase the apparent direct correlation between

[1] From a Master's thesis by H. W. Nutt entitled, "Rhythm in Handwriting." Department of Education, University of Chicago, 1916.

TABLE V. SCATTER DIAGRAM TO SHOW THE EFFECT OF A CONSTANT FACTOR (AGE) IN PRODUCING A SPURIOUS CORRELATION

(The *X*'s represent seven-year-old children, the *O*'s ten-year-old children, and the *V*'s fourteen-year-old children.)

Score in Quality

Rank in Rhythm	20	30	40	50	60	70	80	90	100	110
21		V								
20										
19										
18										
17						V	V			VVV
16					V					
15						V				
14			VV				V			
13				V	VV		V			
12		OV	V	VVV	V		V			
11	O	O		O	O			V		
10	V	O	O	OV	OVV		VV			
9	O	OO		OO		V	VVV			
8	O		OO		O			O		
7			OV			O	V			
6	OO	XO	O			V	V			
5	XXV	XO		O		OV	O			
4	XX	XO	O			O		V		
3	XX	XXX	X		X	X				
2	X	XX	X		X	O				
1	XX XXX	XX XXX	XX	X			X			

quality and rhythm in handwriting. The score of these two characteristics rises as the child grows older. The children who are represented in the diagram are divided into three age groups, of seven, ten, and fourteen years respectively. Those of seven years of age are designated by *X*'s, those of ten years by *O*'s and those of fourteen years by *V*'s. We see that the *X*'s, which represent the younger children, are grouped toward the lower left-hand side of the table. This means that they make low scores in rhythm and also in quality. The *V*'s, representing older children, are grouped toward the upper right-hand corner of the table, showing that they make high scores in both rhythm and quality. The *O*'s fall in between these two groups.

The second fact which is noticeable in the table is that, if we take the symbols representing all of the individuals, without reference to the age distinction, we see that they form a group which is elongated along a diagonal line runing upward and toward the right. This indicates correlation, because it means that those who are low in one test are low in the other, and those who are high in one are high in the other.

If now, in the third place, we look at each of these age groups by itself, we see that the symbols are not grouped along such a diagonal line. The *X*'s alone, or the *O*'s or *V*'s alone, are scattered promiscuously over a given area of the diagram. It is only when we include in our view the three age groups that we find the diagonal grouping to obtain. It is clear, therefore, that there is no marked direct relationship between quality and rhythm, but that, because each one is affected by age, there is an indirect relationship between them when we compare children of various age groups together. An indirect relationship of this sort produces what is called a spurious correlation.

Spearman pointed out that such factors as age may serve

to increase the correlation between two traits or to decrease it. The increase is produced when the common factor affects both traits alike, and a decrease is produced if it affects one and not the other. In addition to pointing out this fact, Spearman presents a formula which is for the purpose of determining what the real correlation would be if the irrelevant factor did not exist. That is, the formula gives a corrected coefficient which is derived from the correlation between the two processes and between each process and the irrelevant factor.[1] His formula has been superseded by a formula for calculating partial correlation — that is, the correlation between two factors which are affected by a third factor, if the third factor is assumed to remain constant.

Spearman's criticism was important because it called attention to the complication of factors which affect a correlation coefficient. It is this fundamental point which is important. It puts us on the watch for spurious factors. When they have been discovered their effect may be eliminated by the application of the partial correlation formula, or it may be avoided by the initial selection of the persons who are to be tested. For example, the disturbing effect of age may be avoided by choosing children who are all of the same age. This is perhaps the preferable procedure when a large enough number of cases can be secured.

Spearman's fourth criticism had to do with the effect of chance factors or errors in measurement or observation upon the correlation coefficient. He pointed out that if the test itself is inaccurate, or if both tests are inaccurate, the apparent coefficient is lower than it would otherwise be. The coefficient which is found by using the inaccurate raw scores he called the *raw coefficient*. He devised and presented a

[1] This description fits the 1904 article. Spearman subsequently modified the formula.

formula for determining what the true relationship would be if the tests themselves were accurate. This he called the *true correlation.* The application of this formula is called *correction for attenuation.* The assumption underlying the correction formula is that the true correlation will be higher than the raw correlation in proportion as the original tests are inaccurate.

Here again Spearman's criticism is probably more important than his correction formula. Certainly, from the point of view of individual diagnosis, it is of no value to know that the correlation would be higher if the tests were more accurate. If the test is inaccurate the ratings of individuals will be unreliable. Whether the formula has a greater theoretical than practical importance may be a debatable question. At any rate, from the practical point of view, the chief importance of this criticism is in calling attention to the necessity of so perfecting our tests that they are accurate measures of the trait which they do represent. If a test is found to be unstable, the correct procedure is to perfect it until it gives as nearly consistent results as it is possible to obtain.

The method of determining accuracy is to give a test twice, give two forms, correlate half of the test with the other half, or make a similar comparison, and then find the correlation between the two sets of scores, called the *reliability coefficient.* If the coefficient is high, the test may be relied upon for consistent results. If low, the results are inconsistent, and no very sound conclusions can be drawn from the tests either with reference to individuals or with reference to the magnitude of the real correlation.

This last criticism of Spearman's has been influential in leading to a refinement in the methods of giving tests which comprises a considerable part of what we call *standardization.* The necessity of such standardization is represented,

for example, in Whipple's *Manual of Mental and Physical Tests*.[1] In this book, which was first published in 1910, are described in detail the methods of giving some fifty single tests. These methods have been carefully worked out and

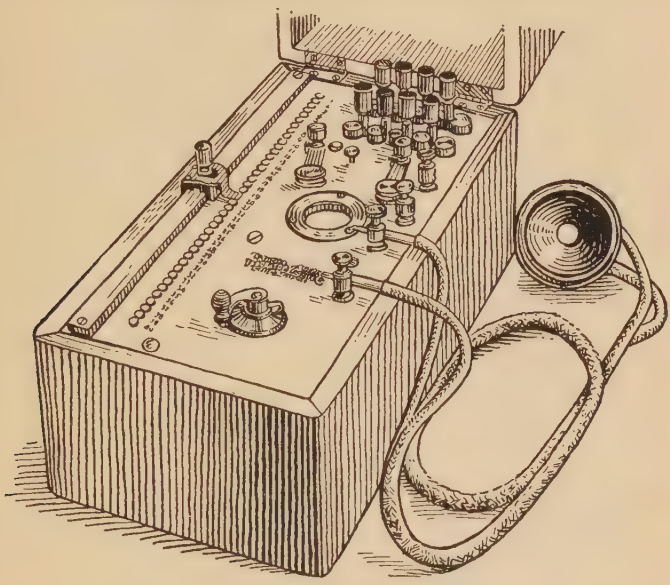


FIG. 1. THE SEASHORE AUDIOMETER

This is an illustration of one of the instruments which represent the standardization of tests. Many such instruments are described by Whipple. (Courtesy of C. H. Stoelting Co.)

are minutely described. The purpose of this standardization of methods of giving tests is to make the score constant — to insure that, on successive tests, it will be a measure of the same thing.

2. Correlation studies of single tests

The critical discussion of Spearman stimulated a number of intensive and elaborate studies of tests, particularly with

[1] Guy Montrose Whipple, *Manual of Mental and Physical Tests*. Baltimore: Warwick & York, Inc., 1924.

reference to correlation. The first, and one of the most careful of these, was made by Cyril Burt,[1] in 1909. We may take this as typical of the studies on this subject.

Burt gave a series of twelve tests to two groups of boys, all of them about twelve years of age. He also secured from the teachers an estimate of the general mental capacity or intelligence of the boys. One of the groups was composed of students in a superior elementary school in England, and the other in a preparatory school. Burt's tests included a number which were characteristic of the earlier period of tests, namely sensory discrimination and motor ability. He also gave two sensori-motor ability tests, which required more complex response than the motor tests, several tests of association, and one which he called a test of voluntary attention.

These tests, classified according to Burt's description, are as follows:

Sensory discrimination:
(1) Esthesiometer test — discrimination of two points on skin.
(2) Weight discrimination.
(3) Pitch discrimination.
(4) Discrimination of length of lines.
Motor tests:
(5) Tapping.
(6) Dealing cards.
(7) Card-sorting.
(8) Alphabet-finding.
Association tests:
(9) Immediate retention.
(10) Mirror drawing.
(11) Spot pattern.
Voluntary attention:
(12) Dotting.

[1] Cyril Burt, "Experimental Tests of General Intelligence," *British Journal of Psychology*, III (1909), 94–177.

The card-sorting tests differed from the dealing tests in this respect. In dealing cards they were put into piles without regard to the character of the cards themselves. In sorting, they were sorted according to the numbers on the cards. This involved a discriminative response each time a card was thrown into a pile. The alphabet-finding test was carried out in this way: A number of cards, each containing one letter of the alphabet, were placed before the subject in an irregular arrangement. The task was to pick out the letters in the alphabetical order and arrange them beneath the original group. The mirror-drawing test required the subject to draw a figure which was seen in a mirror and therefore reversed. In the spot-pattern test the experimenter showed, for a very short time, a figure marked off into squares. At various intersections of this figure were dots. The individual was then given a card which had similar lines upon which he was required to place dots in the same position as in the card shown him. In the dotting test the individual being tested was seated before a rotating disk, covered, except for a narrow slot, by a card. On the disk were placed dots in irregular positions. The individual was required to strike these dots with a pencil as they appeared one at a time through the slit. This required alertness and a rapid adjustment of the movement of the hand to the position of the dot.

Burt worked out the technique of the administration of each of these tests with considerable care, and, in order to determine how consistent or reliable the scores of each test were, he gave each one twice, and then calculated the correlation between the two sets of scores for each group. These reliability coefficients varied considerably. The lowest was .38, which was found in the card-sorting test in the preparatory-school group. The highest was .93, in the test for memory of words and syllables, also in the prepara-

tory group. Eleven of the twenty-two reliability coefficients which were reported were below .70, and the other eleven were .70 or above. It will be remembered that a coefficient which expresses positive correlation may vary from zero to one. A correlation which approaches one therefore, is high, and one which approaches zero is low.

It is a matter of judgment as to how high a reliability coefficient must be in order that the test may be regarded as satisfactory. Certainly, a coefficient as low as .50 represents little reliability. Perhaps we may say that .70 is the lower permissible limit of such a coefficient. Anything below this means that the test gives very variable results when repeated, and one can place but little reliance upon a single score even with a correlation of .70. A reliability coefficient of .90 would now be considered necessary. We see, then, that about half of Burt's tests met the very liberal requirements mentioned so far as reliability is concerned.

Burt used these reliability coefficients also to calculate the true scores according to Spearman's formula, which has already been mentioned. We shall not be concerned with these, but shall use only his raw scores.

The next question we may ask concerning Burt's result is, which of the tests gave high correlations with the other tests in general, and which ones gave low correlations? The results agree with the conclusions which we have already reached from the earlier experiments. The tests of discrimination and of motor ability have little relationship to the other tests, or to each other. The tests which Burt designates as measures of association or of voluntary attention are the ones which have the highest correlation among themselves, and with other tests in general. The order in which the tests fall, as measured by the degree with which they correlate with other tests, is as follows:

Elementary School	Preparatory School
Dotting test	Dotting test
Alphabet test	Alphabet test
Card-sorting	Mirror test
Card-dealing	Memory
Spot pattern	Spot pattern
Tapping	Tapping
Mirror	Sorting cards
Pitch discrimination	Pitch discrimination
Discrimination of lines	Discrimination of lines
Touch discrimination	Weight discrimination
Memory	Touch discrimination
Discrimination of weight	Card-dealing

In a few cases the order in which the particular tests fall differs rather widely in the two lists. This is true notably of the card-sorting, the card-dealing, the mirror test, and the memory test. On the whole, however, the two lists correspond very well, which means that the tests which show a high correlation with other tests in one group also show a high correlation in the other group. The correlation between the order of the two lists of tests is .80.

We may now examine the tests from a point of view of the extent to which they correlate with estimates of intelligence made by the teachers, or imputed intelligence, as Burt calls it. The order in which the tests fall, based upon the closeness of their correlation with imputed intelligence, is as follows:

Elementary School	Preparatory School
Spot pattern	Dotting
Mirror	Alphabet
Alphabet	Memory
Dotting	Spot pattern
Memory	Sorting
Sorting	Mirror
Tapping	Tapping
Dealing	Pitch discrimination
Pitch discrimination	Dealing
Discrimination of lines	Discrimination of lines
Touch discrimination	Touch discrimination
Weight discrimination	Weight discrimination

There is a rather close agreement between the order of the tests in the two schools as judged by this criterion, and also between the order based on the intercorrelation between tests and that based on the correlation with imputed intelligence. In other words, the tests which agree closely with the estimates by the teachers of the pupils' abilities also agree closely, in general, with the other tests.

This method of rating a test on the basis of the closeness of its agreement with other tests, or with imputed intelligence, suggests a more elaborate comparison and an interpretation of the constitution of mental ability. This more elaborate comparison is made on the basis of a table which shows the correlation of every test with every other test. For brevity, we shall study only the table which was derived from the results of the elementary-school group, Table VI.

Table VI is arranged in this fashion. Every test is represented by a horizontal row of coefficients and by a vertical column of coefficients. Thus, the top row and the first column contain the coefficients from the dotting test. The next row and the next column represent the alphabet test. The fourth row and the fourth column represent the results from the correlation of imputed intelligence. The table is so constructed that there is a place for the coefficient of correlation between every test and every other test. By following each horizontal row to the point of its intersection with a given vertical column we may find the correlation between the tests which are represented in the given row and column. One coefficient in each row represents the correlation of a test with itself, or the reliability coefficient. The reliability coefficients are enclosed in heavy lines.

Tables of the intercorrelation of mental tests of this sort form the basis of much discussion concerning the nature and the relationship of mental capacities. Spearman and Krueger, in their earlier studies, presented tables of this sort,

but they were not extensive enough to serve as satisfactory evidence upon the problem. Burt was the first investigator to furnish adequate material for the construction of such a table.

TABLE VI. THE INTERCORRELATIONS OF THE TESTS GIVEN BY BURT TO THE BOYS OF THE ELEMENTARY SCHOOL

(The numbers at the head of the columns stand for the same tests as are named after these numbers on the left side.)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Dotting	.86	.77	.67	.60	.69	.57	.57	.50	.52	.48	.38	.20	.16
2. Alphabet-finding	.77	.60	.74	.61	.66	.59	.53	.20	.52	.16	.62	.01	.07
3. Card-sorting	.67	.74	.84	.52	.72	.45	.61	.34	.52	.14	.22	.19	.23
4. Imputed intelligence	.60	.61	.52	.88	.44	.76	.47	.07	.40	.29	.13	.57	—.13
5. Card-dealing	.69	.66	.72	.44	.88	.51	.65	.40	.34	.47	.23	.19	—.13
6. Spot pattern	.57	.59	.45	.76	.51	.55	.41	.45	.47	.25	.03	.26	.11
7. Tapping	.57	.53	.61	.47	.65	.41	.51	.45	.47	.08	.26	.05	.22
8. Mirror drawing	.50	.29	.34	.67	.40	.45	.45	.52	.34	.16	.08	—.05	—.05
9. Pitch discrimination	.52	.52	.52	.40	.34	.47	.47	.34	.67	—.07	—.01	.01	—.13
10. Line discrimination	.48	.16	.14	.29	.47	.25	.08	.16	—.07	.50	.26	.06	.19
11. Touch discrimination	.38	.62	.22	.13	.23	.03	.26	.08	—.01	.26	.73	.16	.29
12. Memory	.20	.31	.19	.57	.19	.26	—.05	.05	.01	.06	.16	.70	.05
13. Weight discrimination	.16	07	.23	—.13	.01	.11	.22	—.05	—.13	.19	.29	.05	.86

The order of the tests is based upon the average degree of correlation between the individual tests and all of the others. The test having the highest degree of correlation is at the top and the one having the lowest is at the bottom. The fundamental observation which is made concerning a table of this sort by Spearman and Burt is that there is a

kind of consistency about the relationship between the mental traits which is significant. This consistency is shown by the fact that those tests which have high correlation on the average have relatively high correlation with each of the individual tests. The dotting test, for example, has a higher correlation with the alphabet-finding test than has any of the others. Its correlation with the card-sorting test is higher than that of most of the other tests. In similar fashion, it correlates to a high degree with imputed intelligence, with card-dealing, with the spot-pattern test, and so on. This rule has exceptions, but we shall pass them over for the moment.

3. Factor analysis

This consistency in the degree of intercorrelation between the various tests is expressed in the term *the hierarchy of intelligences*. By hierarchy of intelligences is meant that situation in which some mental traits are superior to others as measured by the degree of their correlation with other mental traits. They stand higher on the scale as measured by intercorrelation. Other traits, on the other hand, stand lower on the scale as measured in this same fashion. It is evident that if the rank order of the coefficients which represent the intercorrelations of the various tests with one particular test is approximately the same as that of these same tests with a second test the columns and the rows of the coefficients in a table of intercorrelations will themselves be positively correlated. This correlation between the magnitude of coefficients in such a table has been taken as a measure of the extent to which the hierarchical arrangement obtains.

This fact that a hierarchy exists, so that some capacities are correlated relatively closely with all other capacities while other capacities have in general only slight correla-

tion, has led Spearman and his school to set up the hypothesis of a central factor which is shared in large measure by some capacities or traits and in small measure by others. This central factor is regarded as responsible for the correlation between tests. It was originally called general intelligence, but is now called by Spearman merely "*g*," so as not to confuse it with other meanings of intelligence. A fuller discussion of the evidence for "*g*" will be presented in the chapter on "The Nature of Ability."

We may now follow up the theory of this school a little further.

In his first study of the hierarchical arrangement of coefficients of intercorrelation between tests, Spearman took as a measure of the approach to a perfect hierarchy the degree of correlation between the series of correlation coefficients which occupy the several columns of the table.[1] Take, for example, columns one and two, Table VI, p. 77, which represent the dotting test and the alphabet test. The assumption of the hierarchy of intelligences is that if a particular test — say the card-sorting test — correlates highly with the dotting test in column one, it will also correlate highly with the alphabet-finding test in column two. On the other hand, if another test, say the memory test, correlates to a low degree with the dotting, it will also correlate to a low degree with the alphabet-finding. Now, if this rule holds throughout, there will be a high correlation between the coefficients of column one and column two. Similarly, there will be a high correlation between all of the other columns. Further, if the tests are arranged in the descending order of their average correlation with all the tests, they will also be in a descending order with respect to the correlations with each individual test, as has already been said. Hart and

[1] B. Hart and C. Spearman, "General Ability, Its Existence and Nature," *British Journal of Psychology*, V (1912), 51–84.

Spearman find the correlation between columns of tables of this sort to be in general fairly high. They range from .58 to .98. This they consider to have a very significant bearing upon the constitution and relationship of mental ability.

The matter may be simplified by considering only four abilities, for example, the first four in Table VI. The intercorrelations are as follows:

	3 Card-Sorting	4 Imputed Intelligence
1. Dotting	.67	.60
2. Alphabet-finding	.74	.61

Test 1 correlates more closely with other tests in general than does test 2. It should then have a higher correlation with a particular test than should test 2. That is, the correlation between 1 and 3 should be higher than the correlation of 2 with 3. Similarly the correlation of 1 with 4 should be higher than the correlation of 2 with 4. Moreover, the relations of these correlations should be proportional, thus:[1]

$$\frac{r_{13}}{r_{23}} = \frac{r_{14}}{r_{24}}$$

The reason for this proportionality and the explanation of its existence are found in the two-factor theory, which will be described in a moment.

The above equation may be expressed thus:

$$r_{13} \cdot r_{24} - r_{14} \cdot r_{23} = 0$$

The differences given by this formula are called tetrad

[1] C. Spearman, "Some Issues in the Theory of 'G' (Including the Law of Diminishing Returns)," *Proceedings, British Association*, Section J. Southampton: 1925.

differences. In the example drawn from Burt the tetrad difference would be:

$$.67 \times .61 - .60 \times .74 = -.0353$$

Apparently the relations are not proportional, as the theory would presuppose, but the use of this single example assumes that every coefficient is a true measure of the correlation in question. In other words, it neglects the error of sampling.[1] We should not expect the tetrad differences as worked out by the above formula to be all zero. We should expect their average to be zero, and that their probable error should be that which is to be expected from the theoretical and the actual probable errors of the tetrad differences. Spearman worked out the theoretical and actual probable errors for Simpson's table of intercorrelations and found them to be respectively .061 and .062. From another table of intercorrelations between measures of physical traits, which we should not expect to be arranged in the form of a hierarchy, he found the actual and theoretical probable errors to be respectively .089 and .011.

The early discussion of correlation and Spearman's tetrad equation have been presented in some detail because of the historical importance of Spearman's work. For many years Spearman's method, and his theory of abilities which emerged from it, was the only method of factor analysis and the only theory of abilities attached to it in the field. Not all writers on the subject agreed with his interpretation but no others furnished an alternative analysis.

In the last dozen years several forms of analysis which are different in detail but based on the same fundamental principle and interchangeable mathematically have appeared. Two of these have the most direct bearing on the

[1] C. Spearman and K. Holzinger, "The Sampling Error in the Theory of the Two Factors," *British Journal of Psychology*, XV (1924), 17–19.

design of tests, the methods of Thurstone [1] and of Kelley.[2] The bearing of the interpretations which Thurstone and Kelley make of their results on the design of tests and the theory of ability will be brought out at the appropriate points. It is with these bearings that this book is principally concerned. The procedure of factor analysis has become so technical and, in part, so controversial that it cannot be adequately treated in a general text. An exposition which emphasizes the two-factor method of Spearman, amplified by Holzinger to include group factors and called the bi-factor method, is given by Holzinger [3] and one with emphasis on the method of vector analysis of Thurstone is given by Guilford.[4]

This discussion of factor analysis may be concluded by a few general remarks regarding its implication for the theory of abilities and the design of tests. The most important fact to be kept in mind is that the use of factor analysis does not compel one to find one particular set of factors and no others in a set of variables, as, for instance, test scores. For example, a given set of correlations can be accounted for by supposing that there exists one general factor present in all the tests; or a special factor, different in each test; or a small number of group factors each present in part of the tests. On the other hand, the general factor may be dispensed with, and the intercorrelations may be accounted for by the

[1] L. L. Thurstone, *The Vectors of Mind.* Chicago: University of Chicago Press, 1935.

[2] Truman L. Kelley, *Crossroads in the Mind of Man.* Stanford University, California: Stanford University Press, 1928; Truman L. Kelley, *Essential Traits of Mental Life.* Harvard Studies in Education, Vol. XXVI. Cambridge, Massachusetts: Harvard University Press, 1935.

[3] Karl J. Holzinger, *Student Manual of Factor Analysis.* Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1937.

[4] J. P. Guilford, *Psychometric Methods.* New York: McGraw-Hill Book Co., Inc., 1936.

group factors or primary abilities. Not only may these two types be assumed, but a great variety of particular solutions may be found under each type. A third type is also possible, in which the correlations are explained by the overlappings of groups of many particular and narrow factors. The solutions, to use Holzinger's term, are permissive. They are not mandatory.

What then determines which solution one shall choose? One consideration is that of parsimony, which dictates the choice of the simplest possible solution, with as few factors as possible. Why is a simple explanation better than a complicated one? Because it is easier to understand and we like it better. But is it truer to the facts? Not necessarily. The "law" of parsimony is more a tendency of thought than a law of fact. Other things being equal we should doubtless choose the most parsimonious solution, but other things may not be equal.

What else guides the factor analyst? The psychological meaning of the suggested factors. A mathematician might devise factors that had nothing but mathematical significance. A psychologist tries to find factors which he can describe in terms of psychology or common sense. Does this merely mean that he takes out what he puts in? No, but it may mean that he is limited in his discovery of factors by his ordinary psychological concepts.

The existence of factors which are conceived to be uniform from person to person, and from time to time or activity to activity in the same person, is doubted by some psychologists. It appears to be too rigid a conception, not to fit the great variety and multiplicity of human behavior and attainment, nor to describe the fluidity of human activity. It seems to posit a group of entities behind human behavior, like the "faculties" of sainted memory. Perhaps this view is a misconception, and factors can be thought of

as operational concepts, defined by the measures by which they are arrived at. But this is rendered somewhat difficult by the fact that the tests of factors are worked out after the factors have been laid out.

It would seem that some real basis for the choice of factors might be found in the nature of the organism if we knew enough about the organism to give us a basis. An attempt to apply our present knowledge would be largely speculative. We may, perhaps, get hints here and there to assist in deciding particular questions. This discussion, however, is carrying us beyond the subject of factor analysis itself. We shall refer again to some of these questions in discussing the nature of ability.

Chapter IV

AGE SCALES: THE BINET SCALES AND THEIR REVISIONS

The studies of correlation between single tests, which have been described in the last chapter, did not at once issue in the development of practicable scales. They were used later in the development of our so-called *point scales*, and this application will be described in connection with the account of these scales. The first successful scales of intelligence were those of the type which Binet developed.

1. Binet's early experimental work

As we have already seen, Binet, in the earlier period, experimented with tests of the type which had been used by other psychologists. These tests measured the simpler mental capacities separately. His age scale differs from most of his previous tests in that it includes many tests of the higher or more complex mental processes, and in that it was so arranged that the scores which the pupil makes in the various component tests of the scale can be combined into a composite score.

The idea which led to the use of a composite score to express the total results of the pupil's reaction to the tests of the scale was the idea of mental age. Binet apparently approached this idea in the beginning in a somewhat indirect fashion. The first scale which he put out, in 1905, was simply a series of tests of widely different degrees of difficulty, arranged in order from the easiest to the hardest. Such a series of tests is very appropriately called a scale, because it ranges upward in difficulty.

We may pause to consider briefly some of the character-

istics of this 1905 scale.[1] It consisted of thirty tests, some of them being composed of several parts. One of these, in fact, included twenty-five individual tests. This multiplicity of tests is the first feature of significance, and one which is largely responsible for the success of our so-called *intelligence tests*. It has been found that, while single tests may sometimes have fairly good reliability and validity, groups of tests are much superior in these respects.

As has been said, the tests range from very easy to relatively difficult. The first test, in fact, is much easier than the easiest one in Binet's later scale. It requires that the child should follow with his eye the course of a lighted match which is passed in front of his face. The fifth test of the series represents a higher stage of development. The examiner wraps a piece of candy in a paper in the sight of the child and hands it to him to see if he removes the paper and eats the candy. In the tenth test the child is shown two lines and is required to tell which line is the longer. In test fourteen he is asked to give the meaning of the words *spoon*, *house*, *dog*, *mamma*. In test sixteen he is asked to tell the difference between well-known objects, such as paper and cloth. In test twenty-two he is asked to place in order five weights of eighteen, fifteen, twelve, nine, and six grams respectively. In test twenty-six he is asked to put in one sentence the three words *parrot*, *bank*, and *fortune*. In test twenty-seven he is asked to respond to a number of questions. This is the test which contains twenty-five items. Among them are: "What should one do when he is cold when there is a fire in the house?" "What happens when one is lazy and does not wish to work?" "Why is it better to persevere in what one has begun, than to give it up and try

[1] A. Binet and T. Simon, "Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux," *Année psychol.*, t. 11 (1905), 191–244.

something new?" In the thirtieth test the child is required to tell the difference between *esteem* and *friendship*, and between *remorse* and *chagrin*. The second important characteristic of this series of tests, then, is that it is a scale of increasing difficulty. This makes it possible to test abilities of wide range, and to place the individual on a scale of ability which is fairly continuous and uniform from one level of experience to another.

In Binet's comments on this preliminary scale is contained the germ of the later age scale. While the tests themselves are not classified according to years, Binet indicates what a child of three, five, seven, nine, and eleven years may be expected to do on the tests. Thus he tells us that a child of three years names and recognizes the names of the majority of objects which figure in his everyday life. A child of five years, for the most part, repeats three figures, compares two lines, and after a lesson, two weights; he is likewise able to define the names of familiar, concrete objects.

Finally, this 1905 scale involves the central idea according to which a difference in scores on an age scale is interpreted as being a measure of differences in intelligence. Individuals of different degrees of intelligence at the same level of maturity are distinguished by the number of tests they can pass, just as are children of different ages. An idiot, for example, can pass the first six tests, or at least can pass no higher than the first six. The capacities of imbeciles are represented by the range from tests seven to fifteen, and of feeble-minded by the range from sixteen up to the point which represents normal intelligence. These levels of difficulty in the tests, and the levels of ability which they represent, stand for the ultimate mental status of the various types of individuals. The idiot will never develop beyond the stage represented by test six. Imbeciles will never

develop beyond the stage represented by test fifteen, and so on. This means, then, that an adult of a certain degree of mental ability, represented by mental defect of certain amount, reaches ultimately the mental capacity of a normal child of a given age.

We have now reached the fundamental idea of the Binet scale and, in fact, the fundamental idea of all of the scales by means of which the mental capacity of children is measured by interpreting their scores in terms of age norms. This idea is the identification of differences in mental capacity, or in brightness, above and below that of the average person, with differences in stages of mental development as represented by the capacity of children of various ages.

2. The Binet-Simon scale of 1908

The plan of using the stages of mental growth of a normal child as a scale by which to measure differences in intellectual capacity, or differences in brightness, is represented first in the concept of *mental age*, and its use as a measuring unit. This idea has its complete development in the next scale which Binet, in collaboration with Simon, put out in 1908.[1] In the 1908 scale each test is classified under some one age. The ages from three years to thirteen years, inclusive, are represented. The number of tests at each age varies from three at age thirteen, to eight at age seven. Illustrations from two of the years will serve to represent the entire scale.

Five Years

1. Comparison of two weights, one pair three and twelve grams respectively, the other pair, six and fifteen grams respectively.
2. Copying a square with pen and ink.

---

[1] A. Binet and T. Simon, "De Développement de l'intelligence chez les enfants," *Année psychol.*, t. 14 (1908), 1–90.

3. Putting together two triangles so as to make the same form as a rectangle.
4. Counting four pennies.

ELEVEN YEARS

1. Detecting the absurdity of a series of statements.
2. Building a sentence out of three words.
3. Naming any sixty words in three minutes.
4. A definition of abstract terms.
5. Arranging a number of words which have been disarranged so as to make a meaningful sentence.

The method of using the scale, which is essentially the same as that of using the later revision, is this. The examiner gives the child the tests in order of their difficulty, beginning at the point at which the child can probably pass all of the tests. He then proceeds upward until he reaches the point at which the child fails on all of the tests of an age group. He next estimates the mental age of the child by taking as a basic age the point at which the child passes all of the tests. Then he adds to this one year for every five tests which the child passes beyond. This gives the child's mental age.

It is apparent that the mental age thus represents the composite of a child's ability on a considerable number of tests. It is not required that he pass any particular test. If he fails on one, he may make it up by passing another. His mental age, then, represents a kind of average score which corresponds with what the average child of a particular chronological age can do.

It will be apparent on a moment's thought that, while the mental age may represent the child's maturity, it does not directly represent his intelligence or his brightness. If an eight-year-old and a ten-year-old child have the same mental age, they may occupy the same level of maturity, but they do not possess the same degree of brightness. The

brightness or intelligence of a child must be found by taking account of the relationship between his maturity, or his mental age, and his chronological age. Binet did not work out this problem to a final solution. He simply regarded a child as superior if he passed a test a year or two in advance of his chronological age, and as retarded if he passed only the tests a year or two below his chronological age.

3. The 1911 revision

After the appearance of the 1908 scale, a number of psychologists applied it to children and reported upon the results. Among these were Decroly and Degand,[1] of Belgium, who reported that some of the tests were too easy and others too difficult. In Germany, Bobertag [2] gave the tests to a considerable number of children and reported very fully upon the responses which were made to each test of the scale. He reported particularly the percentage of children of each age who passed the various tests. He then attempted to determine where the tests belong on the scale by placing them at that age at which 75 per cent passed. Binet had apparently used such a standard, but Bobertag attempted to apply it in more exact fashion. In the United States, Goddard [3] applied the test to the feeble-minded children of the Training School at Vineland, New Jersey, and also applied it to a large number of normal children in the elementary school.[4] He also criticized the test on the ground that in-

[1] O. Decroly and J. Degand, "La mesure de l'intelligence chez des enfants normaux d'après des tests de Binet et Simon," *Arch. de psychol.*, t. 9 (1909), 81–108.

[2] O. Bobertag, "A. Binet's Arbeiten über die intellectuelle Entwicklung des Schulkindes (1894–1909)," *Zeitsch. für Angew. Psychol.*, B. 3 (1909), 230–59.

[3] H. H. Goddard, "Four Hundred Feeble-Minded Children Classified by the Binet Method," *Pedagogical Seminary*, XVII (Sept., 1910), 387–97.

[4] H. H. Goddard, "Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence," *Pedagogical Seminary*, XVIII (June, 1911), 232–59.

dividual tests were not properly placed, and that certain parts of the scale were too easy and other parts too difficult. However, he reported that the scale as a whole was extremely reliable. This conclusion he drew from the fact that the distribution of the rankings of children taken as an entire group was practically in agreement with the normal distribution curve. His results, and those of other investigators, however, showed that the distribution for some of the individual ages was far from normal. For the lower ages the tests were too easy, and the majority of the scores were above the normal, whereas for the upper ages they were too hard and the bulk of the scores were below the normal.

Taking account of these studies and of the recommendations which were based upon them, Binet revised his scale and published the revised form in 1911.[1] Due to his untimely death, this was Binet's final contribution to the testing movement.

The changes in the scale were of two sorts. In the first place, it was made more uniform by having an equal number of tests for each age, namely, five. This avoided the error which was present in the calculation of the mental age in the 1908 scale, due to the fact that at some parts of the scale it was easier to obtain advanced credit than at other parts. The other type of change was the transposition of some of the tests to different ages. In some cases, because the tests were apparently too difficult, the tests were moved to higher ages. In other cases, because they were too easy, they were transposed to lower ages. In order to overcome the excessive difficulty of the scale at the upper end, the tests for eleven years were moved up to the twelve-year period, and those for twelve years to fifteen years. The thirteen-year-old tests were called adult tests, and two others were added to

[1] A. Binet, "Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école," *Année psychol.*, t. 17 (1911), 145–201.

them. This shift, however, did not meet the difficulty, since tests were not substituted for the years eleven, thirteen, and fourteen.

The method of finding the mental age was changed in one respect, in that the child's basic age was taken at the age at which he passed all but one test, instead of every test. This made the scale somewhat more flexible.

It is not certain whether the 1911 scale was an improvement over the 1908 scale. Some investigators who have compared the results of the two prefer the original scale. The changes which were made were made largely in response to criticisms by others, rather than because Binet's own experience indicated their desirability. The fact that the ground of the changes was not certain is shown by the inconsistency between the changes which were made by Binet and those made by Goddard in his revision, which also appeared in 1911. We may pass to a brief account of this and other revisions made by other workers.

4. Other revisions of the Binet scale; Goddard's 1911 revision

One of the earliest and most enthusiastic users of the Binet scale in the United States was H. H. Goddard, the chief psychologist in the Training School at Vineland.[1] Goddard very early began working with Binet's scale, adapting it to American conditions. In some respects the changes which he made were similar to those which were made by Binet. For example, above the fourth year he provided the same number of tests for each age, namely, five. In general he made fewer changes in the position of the tests than did Binet, and in most cases retained in their original position the tests which had been moved by Binet.

[1] Henry H. Goddard, "A Revision of the Binet Scale," *Training School Bulletin*, VIII (March, 1911), 56–62.

He retained the tests for eleven and twelve years, but moved the thirteen-year tests up. He introduced a number of tests of his own into the fifteen-year age group. Goddard, of course, adapted the terminology and to some extent the content of the tests to the experience of American children. This scale was very widely used in the United States until it was superseded by the more extensive revision made by Terman and his collaborators — the Stanford Revision.

Kuhlmann began working with the Binet scale rather early and has produced two revisions. The first appeared in 1912, and the second in 1922.[1] Since the second revision is a modification and extension of the first, we may confine our discussion to it. The most important contribution which was made in Kuhlmann's revision consists in the extension of the scale at both ends, particularly at the lower end. His tests begin as low as three months and enable the examiner to calculate the mental capacity of very young children. At the upper end they extend to fifteen years. Further changes consist, first, in the standardization of procedure. This was generally found to be necessary before one could satisfactorily use Binet's original scale. A number of American authors have contributed to the scale by defining the method of procedure in giving it. In the second place, the scale was modified in the direction of making it more difficult at the lower end, and easier at the upper end. This was done by changing the location of tests. Again, nineteen of the original tests were eliminated because they were found to be unsatisfactory. To those which remain were added a large number, so that there were in the final scale eight tests for each age group above two years. The

[1] F. Kuhlmann, *A Revision of the Binet-Simon System for Measuring the Intelligence of Children.* Journal of Psycho-Asthenics Monograph Supplement, Vol. I, No. 1, September, 1912; F. Kuhlmann, *A Handbook of Mental Tests.* Baltimore: Warwick & York, Inc., 1922.

total tests in the revised scale are one hundred and twenty-nine. This constitutes a very fundamental and thoroughgoing revision, and the scale would doubtless have had wider use if it had not been anticipated by equally extensive revision made by Terman and his collaborators at Stanford University.

5. The first Stanford Revision of the Binet scale

The Stanford revision of the Binet scale is the most widely used age scale. It was extensively employed in the schools previous to the World War. During the War it was one of the two individual tests which were used with the English-speaking recruits who passed a low score on the group tests. Its purpose was to confirm the standing on the group tests and to provide a more accurate measure of intellectual ability than the group tests afforded. It has continued to be widely used in the schools up to the present.

Preliminary experimentation leading up to the Stanford revision was begun by Terman in collaboration with Childs in 1911 or 1912. The results of the preliminary experimentation were published in 1912.[1] The purpose of this preliminary investigation was to secure the scores from children of different ages on a large number of new tests. These tests were to be used to supplement the tests of the original scale. Some of these new tests were incorporated into the Stanford Revision. We shall call attention to these new tests in the more detailed description of this revision.

6. Description of the first Stanford revised scale

The new scale contained ninety items. Fifty-four of these are from the Binet scale, and thirty-six are new. Twenty-seven of these new tests were devised and standard-

[1] Lewis M. Terman and H. G. Childs, "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal of Educational Psychology*, III (February, March, April, and May, 1912), 61–74, 133–43, 198–208, 277–89.

ized in the previous investigation made by Terman. The other new tests were borrowed from other investigators or adapted from earlier Binet studies.

In the tests which were selected from the Binet scale many changes were made. Eighteen were shifted downward one year, four downward two years, two three years, and one six years. Three were shifted upward one year and one two years. The location of each of the tests was determined by the results of their application to about one thousand children of a community near Stanford.

The completed scale contains tests for each age from three to ten, and in addition tests for ages twelve and fourteen, and for average adult and superior adult intelligence. There are six tests for each age except year twelve, for which eight tests are provided. The credit for each test is so arranged that one year in mental age may be gained by passing the tests for each year. Thus, up to year ten, in which there are six tests for each age, the credit in mental age for passing each test is two months. For year twelve, in which there are eight tests, the credit for each is three months, which gives a total credit of twenty-four months. This stands for both year eleven and year twelve. For year fourteen, there are six tests of four months' credit each, or a total credit of twenty-four months, and so on. For the average adult there are six tests with five months' credit for each, or a total credit of thirty months, and for superior adults six tests with six months each, giving a total credit of thirty-six months. The reason for the extra credit in the case of the average adult and the superior adult groups is that, as one approaches the end of the scale, the possibility of making credit by passing advanced tests is progressively reduced.

As an illustration of the tests of the scale we may take those for age twelve. A copy of the record booklet containing the tests for this age is reproduced on page 96.

A DEFENDS DOG GOOD HIS BRAVELY MASTER

TO ASKED PAPER MY TEACHER CORRECT I MY

FOR THE STARTED AN WE COUNTRY EARLY AT
HOUR

YEAR XII. (8 tests, 3 months each, or 6 tests, 4 months each.)

*1. Vocabulary, 40 words. Score..... Total Vocab.....

2. Abstract words. (3 of 5.)

a. Pity.....

b. Revenge.....

c. Charity.....

d. Envy.....

e. Justice.....

3. Ball and field. (Superior plan.).....

*4. Dissected sentences. (2 of 3. 1 minute each.)

a. ..... Time.....

b. ..... Time.....

c. ..... Time.....

*5. Fables. (Score 4, *i.e.*, two correct or the equivalent in half credits.)

a. Hercules and wagoner.....

b. Maid and eggs.....

c. Fox and crow.....

d. Farmer and stork.....

e. Miller, son and donkey.....

*6. Repeats 5 digits backwards. (1 of 3. Read 1 per second.)

3-1-8-7-9..... 6-9-4-8-2..... 5-2-9-6-1.....

*7. Pictures, interpretation. (3 of 4. "Explain this picture.")

a. Dutch Home.....

b. Canoe.....

c. Post Office.....

d. Colonial Home.....

*8. Gives similarities, three things. (3 of 5. "In what way are —, —, —, alike?")

a. Snake, cow, sparrow.....

b. Book, teacher, newspaper.....

c. Wool, cotton, leather.....

d. Knife-blade, penny, piece of wire.....

e. Rose, potato, tree.....

The vocabulary test was added in the Stanford Revision. It consists of one hundred words selected at random from the dictionary, and arranged from easy to difficult or from familiar to unusual words. These hundred words are considered a sufficiently large sample of the pupil's vocabulary to enable one to calculate roughly his total vocabulary. Experiments indicate that the scores made with different groups of one hundred words similarly chosen at random are only slightly different. The test called "abstract words" requires the defining of the words given in the blank. The definitions do not have to be formally accurate but must express the essential meaning.

The ball-and-field test is given by showing the child a circle with a gap on one side. The circle represents the fence around a field and the gap a gate in the fence. The child is told that he is to imagine that he has lost a ball in the field and is instructed to trace with a pencil the path he would take to find the ball. At age twelve the child must follow a superior plan.

The fourth test requires that the child make a sentence out of each of the groups of words printed at the top of the page. They are printed so that they will be right side up to the child.

The fables test consists in five of Aesop's fables of which the child is to give the lesson or the meaning. For example, he must be able to express the meaning of the fable of the maid and the eggs in some such general statement as "Don't count your chickens before they are hatched." An interpretation in more particular terms is a satisfactory response for a younger child.

The tests in repeating digits backwards and in giving similarities need no explanation. The picture interpretation test requires that the child give an explanation of the meaning of the scene portrayed or the story which it tells.

These examples illustrate the character of the tests for the later ages. They employ language to a high degree because the author found language to be the most convenient medium by which to test the higher mental process at these ages.

These minute directions for giving and scoring the test are typical of all of the tests of the scale. The manual which contains these directions occupies 229 pages of Terman's text, *The Measurement of Intelligence.* It is therefore obvious that a considerable amount of careful study and preliminary practice is necessary before an individual is prepared to give this test in such a way as to obtain valid results.

7. The derivation of the first Stanford Revision

The construction and development of this revision is described in a monograph published by Terman in 1917.[1]

The first step in the standardization of the scale was to take the tests in the original Binet scale and the additional tests which had been tried out in a preliminary experiment, or which had been gathered from other sources, and make them into a trial scale. The basis for the selection of the tests for the trial scale was the percentage of children of various ages who passed the tests, as reported by the investigators who used them. This trial scale was then given to about one thousand children up to the age of fourteen, and to about four hundred adults.

In regard to the selection of the children upon whom the test was standardized, the authors make the following statement: "A plan was then devised for securing subjects who should be as nearly as possible representative of the several ages. The method was to select a school in a community of

[1] L. M. Terman and Others, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore: Warwick & York, Inc., 1917.

average social status, a school attended by all or practically all of the children in the district where it was located. In order to get clear pictures of age differences, the tests were confined to children who were within two months of a birthday. To avoid accidental selection, all the children within two months of a birthday were tested, in whatever grade enrolled (below the high school). Tests of foreign-born children, however, were eliminated in the treatment of results" (p. 11).

Since the directions for giving the tests were to be those in the final scale, they were worked out with the care and minuteness which has already been described. The tests were given by a number of students, all of whom had been carefully trained by Terman. The children's responses were, for the most part, recorded verbatim, in order that the records might be rescored if it was necessary to change the difficulty of the scale.

The next step was to score all of the children according to the method which was provisionally fixed upon, and to find their I.Q.s. The I.Q.s of each age were then thrown into a distribution table. The requirement which was employed in the examination of this table to determine whether or not the scale was correct was that the distribution of I.Q.s of a particular age should be normal, and that the median mental age of the children should correspond to their chronological age. That is, the median child of the ten-year-old group should have a mental age of ten years, and the mental ages of children above and below this mental age should be distributed in such fashion as to form a symmetrical curve of distribution.

As might be expected, the first trial did not come up to this requirement. Adjustments were then made by either changing the location of tests or by changing the standard of scoring, and the children's records were rescored. On the

second scoring and tabulation the scale still failed to come up to the standard. It was only after the third revision and rescoring that it came sufficiently near the standard to satisfy its author.

The care which was taken in the preparation and standardization of the scale has justified itself in all its subsequent use. The intellectual standards which are represented in the scale have been found to be substantially accurate in the large number of cases in which the test has been given to English-speaking American children.

8. Measures which are derived from the scale

In the description of the original Binet scale it was remarked that the score which the child made was expressed in terms of mental age. It was further said that the significance of the mental age was to be gathered from its comparison with the child's chronological age, but that this significance was expressed by Binet only in rough fashion when he said that if the child were one or two years below his chronological age mentally, he was to be regarded as retarded.

It was soon discovered by the users of any form of the Binet scale that the significance of one year's retardation or acceleration was different at the lower ages and at the higher ages. One year's retardation was found to be more serious at the lower ages. To put it in another way, approximately twice as many children would be one year retarded at twelve years of age as at six years of age. Or, to put it in still another way, the same number of children were found to be retarded one year at six years of age as are retarded two years at twelve years of age. This means that a given amount of retardation, as expressed in years of mental age, is a variable quantity and depends upon the age of the child.

The possible explanations of this fact will be considered in

the chapter on the technique of mental tests. What concerns us at this time is the fact and its practical meaning in mental measurement. The first man to suggest a measure which should avoid the difficulty which has been mentioned, and which would have the same significance for children of different ages, was William Stern. Stern called his measure the mental quotient.[1] The mental quotient was to be found by dividing the child's mental age by his chronological age. Thus a child whose mental age was equal to his chronological age would have a mental quotient of 1. A twelve-year-old child whose mental age was ten would have a mental quotient of .833, or ten twelfths, while a child whose chronological age was ten and mental age was twelve would have a mental quotient of 1.20.

Terman's statistics convinced him that this type of measure was substantially correct. He called it, however, the intelligence quotient or I.Q., and expressed the quotients as whole numbers by multiplying them by 100. The test of the correctness of the measure is the comparison of the range of intelligence quotients for successive life ages. Terman found that by beginning with the lower ages and comparing successive two-year groups, the middle half of the intelligence quotients for each group covered substantially the same range. The lowest range of the middle fifty per cent of the intelligence quotients was fifteen points, and the highest range seventeen points. Thus the intelligence quotients of the children of five and six years of age covered a range from 97 to 111, while those of eleven and twelve years combined had a range from 92 to 108 (p. 40).

The intelligence quotient, expressed in words, then, means the relation between the child's mental development and what we should expect of him at his age. If a child main-

[1] William Stern, *The Psychological Methods of Testing Intelligence*, p. 80. Baltimore: Warwick & York, Inc., 1914.

tains the same relative intellectual capacity from year to year, he should have the same intelligence quotient. This would not be true, as we have found, of the difference between his chronological and mental age. This difference would increase as he grows older. We have thus in the intelligence quotient a measure which is constant, and by means of which we can compare children of different ages. Assuming that differences in intellectual maturity constitute a means of measuring differences in intellectual capacity or brightness, the intelligence quotient, or the I.Q., becomes a measure of brightness.

The significance of the various intelligence quotients can best be grasped by an examination of the following table, which indicates the percentage of persons who are awarded the various intelligence quotients.

TABLE VII. THE DISTRIBUTION OF INTELLIGENCE QUOTIENTS

The percentage of individuals who have various intelligence quotients or lower:

I.Q.	70	73	76	78	85	88	91	92	95
Per cent	1	2	3	5	10	15	20	25	33.3

The percentage of individuals who have various intelligence quotients or higher:

I.Q.	106	108	110	113	116	122	125	128	130
Per cent	33.3	25	20	15	10	5	3	2	1

This table is to be read as follows: The lowest one per cent of persons in general have an I.Q. of 70 or below. The lowest two per cent have an I.Q. of 73 or below. The lowest twenty-five per cent have an I.Q. of 92 or below. The highest one per cent have an I.Q. of 130 or above. The highest twenty per cent have an I.Q. of 110 or above, and the highest thirty-three and one third per cent have an I.Q. of 106 or above. We see that according to this table the middle fifty per cent of all individuals have intelligence quotients ranging from 92 to 108.

Another way of indicating the significance of the intelligence quotient is to use descriptive terms, thus:

Class	Range of I.Q.s	Class	Range of I.Q.s
Near genius	140+	Border-line	70–80
Very superior	120–140	Moron	50–70
Superior	110–120	Imbecile	25–50
Normal	90–110	Idiot	0–25
Dull	80– 90		

The lower three groups, including the intelligence quotients from 0 to 70, are designated as feeble-minded. Thus a ten-year-old child whose mental age was seven years or less would be classed as feeble-minded, according to this scale.

9. The new Revised Stanford-Binet tests of intelligence

About ten years after the first Stanford revision appeared a second revision was begun. The preparation of this new revision has occupied another ten years. This indicates something of the labor involved in making a mental-age scale. The purpose in making the new revision was to overcome certain faults and limitations of the first revision. This scale was inadequate below the age of four years and above the age of ten. The new scale extends to two years at the lower end and provides a more reliable measure at the adolescent and adult ages. The old scale contained a number of unsatisfactory tests. In some cases the instructions were inadequate. Finally, it contained only one form.

The new scale is designed to overcome these faults. It provides two forms equivalent in difficulty, range, reliability, and validity. Each scale is longer than the old one, containing 129 instead of ninety tests. Tests are provided for half-year intervals below five years and for the years eleven and thirteen. The meagerness of the tests for older adolescents and adults, producing a limitation to the maximum

I.Q.s at these levels, has been corrected by the addition of two supplementary superior adult levels. More performance tests, or tests which use actual objects, have been provided at the earlier levels, but language remains the chief medium at the upper levels.

The standardization was based on more extensive and careful sampling than in the case of the first scale. One hundred children were tested at each half year below six years, two hundred at each age from six to fourteen inclusive, and one hundred at each age from fifteen to eighteen. Tests were given in seventeen communities in eleven states, chosen to represent the East, South, Middle West, and West. Rural and urban populations were sampled, and the various occupational groups were represented in roughly the same proportions as in the population at large. An effort was made to secure samples of the various ages that would not be distorted on account of disproportionate representation in the school grades. Only American-born white children were included.

The tests were chosen for inclusion in the scale on the basis of validity, ease and objectivity of scoring, and such considerations as economy of time, interest, and need for variety. Validity was judged by two criteria, increase in per cents passing from one age to another and "a weight based on the ratio of the difference to the standard error of the difference between mean age (or mental age) of subjects passing the test and of subjects failing it" (page 9). Thus, increase in performance with age is the fundamental criterion.

The rearrangement of the scale was carried on until the mean I.Q. at each age level was approximately 100 (actually a little above 100) and the standard deviation approximately the same at each age. Six revisions were necessary for the first form (Form L), but Form M was produced by matching test for test with Form L.

The authors retain the I.Q. as the measure of the individual's relative intelligence. They discuss the advantages of the standard score, which uses the standard deviation as a unit by which to express the individual's score, but regard the I.Q. as better because it is more widely known and its significance better understood. It may be true that the I.Q. is more convenient, but it is a question whether its inherent ambiguity does not make it better policy to adopt the statistically superior standard score and to educate teachers to understand and use it. The standard deviation of I.Q.s of the revised scale is reported to be approximately 17 points, as compared with about 12 points on the old scale. This difference may be due to the properties of the scale itself or to the variability of the population on whom the test was standardized. Terman [1] is of the opinion that the difference is due entirely to the fact that the new scale was standardized on a greater variety of groups in respect to geographical location, residence in city or country, and economic level. He points to the fact that the old scale yielded a larger standard deviation on some populations than on the population on which it was standardized, as for example, in the survey of an age group by the Scottish Council for Research, in which the standard deviation was found to be 17 for boys and 16 for girls.[2] This fact gives support to the view that the larger standard deviation is due to the more representative character of the population on which the test was standardized.

Critics who have used both the old and the new Stanford Revision agree that the new scale is a more serviceable test of its kind than the old one. It will doubtless have a similarly wide use.

[1] As expressed in a personal letter to the author.

[2] *The Intelligence of Scottish Children*, p. 105. Publications of the Scottish Council for Research in Education. London: University of London Press, 1933.

The broader evaluation of the scale rests on its general plan rather than on the skill with which the plan is carried out. In general, the new scale is based on the same principles as the old one. It is a composite of a variety of different tests. It measures differences in brightness within an age group in terms of differences manifested at various maturity levels. It utilizes the interview method with little emphasis on speed. It is not based on an attempt to define intelligence precisely but uses mental age as an indication of intelligence.

Two general questions may be raised concerning the general plan. First, is the mental-age type of test preferable to the point-scale type? Second, is the composite test preferable to the analytical scale?

Terman argues strongly for the mental-age test, largely on the ground that the significance of the responses may be grasped by the examiner as they are made. The fundamental difficulty with such a scale, however, is the fact that its standardization is laborious, rigid, and final. In a point scale, tests can be added or subtracted and norms may be revised without laborious restandardization of the whole scale. Because of this fact, the new Stanford Revision is probably the last of the mental-age scales.

Whether a composite measure of ability will continue to be widely used depends on the outcome of the present debate concerning the nature and composition of ability. It is the writer's belief that a composite measure, or a measure of general ability, will continue to be used, but that composite tests in the future will be organized so that the score may be analyzed to indicate particular abilities, such as those which are designated primary abilities or group factors, as well as general ability. Tests of undifferentiated ability will doubtless be used for a long time and may continue to be used indefinitely in combination with analytical tests.

DESCRIPTIVE ACCOUNTS IN ENGLISH OF THE BINET SCALE AND ITS IMPORTANT REVISIONS

Goddard, H. H. "A Revision of the Binet Scale," *Training School Bulletin*, VIII (1911), 56–62.

Huey, Edmund B. *A Syllabus for the Clinical Examination of Children with the Revised Binet-Simon Scale for the Measurement of Intelligence.* Baltimore: Warwick & York, Inc., 1912.
Huey's revision.

Kite, E. S. *The Development of Intelligence in Children.* Vineland: The Training School, 1916.
A translation of five articles by Binet and Simon which contain discussions of the methods of intelligence-testing and scale-making, and a description of the three scales of 1905, 1908, and 1911.

Kuhlmann, F. *A Handbook of Mental Tests.* Baltimore: Warwick & York, Inc., 1932.
Kuhlmann's later, more thoroughgoing revision and extension.

Kuhlmann, F. "A Revision of the Binet-Simon System for Measuring the Intelligence of Children," *Journal of Psycho-Asthenics Monograph Supplements*, Vol. I, No. 1, September, 1912.
Kuhlmann's first revision.

Melville, Norbert J. *Standard Method of Testing Juvenile Mentality by the Binet-Simon Scale.* Philadelphia: J. B. Lippincott Co., 1917.

Terman, Lewis M. *The Measurement of Intelligence.* Boston: Houghton Mifflin Co., 1916.
The complete manual for the use of the Stanford Revision.

Terman, Lewis M., and Childs, H. G. "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal of Educational Psychology*, III (February, March, April, and May, 1912), 61–74, 133–43, 198–208, 277–89.

Terman, Lewis M., and Merrill, Maud A. *Measuring Intelligence.* Boston: Houghton Mifflin Co., 1937.
A manual for the administration of the second Stanford Revision of the Binet-Simon scale.

Terman, Lewis M., and Others. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore: Warwick & York, Inc., 1917.
A full account of the procedure followed in making the Stanford Revision.

Chapter V

THE EARLY DEVELOPMENT OF POINT SCALES

1. The first point scale

While the development of point scales, as has already been said, was very largely influenced by the studies of correlation, the first point scale was a direct outgrowth of the Binet age scale. This was the scale which was developed by Yerkes, in association with Bridges and Hardwick.[1] The scale is composed of twenty tests, nineteen of which are taken from the Binet scale. The actual number of tests is somewhat greater than is indicated by this statement, since each one is composed of a number of parts. Thus the test of memory span for digits is composed of ten parts, or five pairs of increasing length and difficulty. The first pair contains three digits each and the last, or most difficult pair, seven digits. In the same way each test contains a short, graded series. In general the easier tests are in the first part of the scale and the more difficult ones in the later part, but there is not a regular gradation in the difficulty of the successive tests. There is, rather, a gradation in the difficulty of the parts within each test.

The tests are not arranged according to age, as we have seen, nor are they scored in terms of age. The various forms of the Binet scale are scored by giving credit for passing each test consisting of a fraction of a year's mental age. Thus, in the Stanford Revision, two months, or one sixth of a year's credit, is given for passing each test. In the point scale, on the other hand, the child is not given credit directly

[1] Robert M. Yerkes, James W. Bridges, and Rose S. Hardwick, *A Point Scale for Measuring Mental Ability*. Baltimore: Warwick & York, 1915.

in terms of mental age, but in terms of points. Thus for each part of the test in memory span for digits the child is given one point. The possible score for this test is five points.

This method of scoring seems to constitute rather a difference in method than in principle, for the point scores are interpreted by comparing them with a table of age standards. Thus, if a child makes a point score of 58 out of a possible 100, his mental age is ten. If his score is 70, his mental age is twelve. The authors of the point scale criticized the Binet scale because it assumes that each stage of mental development corresponds to a certain critical age, and that there is a "correlation between the different functions at different stages of development." It seems that any scale which interprets the scores in terms of age standards assumes this correspondence in the same fashion as does an age scale, and this is true of all of our point scales which are designed for children. After the child's score has been referred to the table of age norms, his brightness score may be calculated in a slightly different way from that which is done with the Binet scale, as we shall see in a moment, but the fundamental conception is the same.[1]

Another point which the authors make in favor of the point scale is that it uses the method of partial credits in scoring, as distinguished from the all-or-none method. This point has some justification, but the partial-credit method can also be applied, and is applied, in a measure, in the age scale. By the all-or-none method is meant that method of scoring in which the child is either given full credit or no credit at all in the test. By the partial-credit method is

[1] For a fuller discussion of the relation between point scales and the Binet scale, see Frank N. Freeman, "A Critique of the Yerkes-Bridges-Hardwick Comparison of the Binet-Simon and Point Scales," *Psychological Review*, XXIV (November, 1917), 484.

meant one in which the child is given some credit if he passes a part of the test, and an additional credit if he passes another part. This is illustrated in the memory test which has already been alluded to. Now it happens that, in this memory test, partial credit is also allowed in the Binet scale, although by a different procedure. Thus, in the first Stanford Revision, a child is given credit for passing one test at mental age 3 if he repeats 3 digits, one at mental age 4 if he repeats 4 digits, at mental age 7 for 5 digits, 10 for 6 digits, 14 for 7 digits, and 18 for 8 digits. The same sort of distribution of graded items of the test at different mental ages appears repeatedly in the age scales.

While the fundamental principle of the point scale, therefore, is not to be found in its difference from the Binet scale, there are some characteristics which commend it from the point of view of convenience. It is easier to revise the norms of the point scale, and it is not essential that every test be given a separate age standardization. A tentative series of age standards may be derived from the application of the test to a small number of children, and then, after the tests have been applied to a larger number, the age standards may be changed, if necessary, in accordance with the accumulation of scores. Furthermore, if it is desirable to do so, it is possible to have different norms for different groups, such as groups belonging to various races or different social environments. It is possible to do the same thing with the age scale by applying a correction to the I.Q., but it involves a clumsier procedure.

The point-scale method of organization is also easier to apply in the development of scales for the measurement of other kinds of mental capacity, such as feeling, or will, or moral attitude. In general it is a more flexible type of organization than the age scale, and is the one which now prevails.

Finally, the point scale has the advantage of being easier to administer than the age scale. It does not require much study to gain a knowledge of the method of presenting it and of scoring it.

The method of reckoning the relative intellectual capacity or the brightness of the child with the Yerkes Point Scale is somewhat different from that used with the age scale. In the age scale, as will be remembered, the child's brightness is found by finding the ratio of his mental age to his chronological age. This means that the child's performance is compared with the performance of other children of other ages. In the point scale, on the contrary, the child's performance is compared only with that of other children of his own age. This is done by finding the ratio of the child's score to the average score of children of his own age. This ratio is called the *Coefficient of Intelligence*, written as C.I. The C.I., like the I.Q., is 1.00 for the normal or average child, above 1.00 for the superior child, and below 1.00 for the inferior child. Just what the relationship is between the distribution of these two ratios has never been worked out. We cannot assume that an I.Q. of 120 means exactly the same as the C.I. of 1.20. A comparison of the implications of the C.I. and the I.Q. is made in Chapter XI.

The Yerkes Point Scale, like the Stanford Revision of the Binet scale, has had very wide use in public schools for making individual examinations. Both of these tests, furthermore, were used for individual examinations of men of low-grade intelligence in the army. Probably the greatest importance of this scale, however, is its influence on the subsequent development of tests, including the army group test.

The edition of the original book describing the Point Scale is now exhausted and a revision of the book has been published,[1] containing an account of the first scale, the Pre-

[1] Robert M. Yerkes and Josephine Curtis Foster, *A Point Scale for Measuring Mental Ability*. Baltimore: Warwick & York, Inc., 1923.

Adolescent Scale, with minor changes, and also an account of two additional scales, the Adolescent-Adult Scale and the Infant Scale.

2. The Herring Revision

The Herring Revision of the Binet test has about the same relationship to the original from which it was derived as has the Yerkes Point Scale.[1] This is also a point scale in that the child is given a specified number of points for each test which he passes, and his mental age is calculated by comparing the total number of points to his credit with a series of age standards. The scale is made up mostly of tests which are derived from the Binet scale, but these are supplemented by several new tests. There are thirty-eight in all.

The chief novelty in the scale is its mode of organization. The tests, instead of being arranged in a single series, are placed in five groups. The first group may be used alone and constitutes a very brief test. The test may be extended by adding the second group to the first, and so on. If any of the groups beyond the first one are used, a scheme is given according to which one may omit part of the tests to avoid duplication of levels of difficulty. If the child makes a relatively high score on the first group, the earlier and easier tests of the second group are omitted. On the other hand, if he makes a relatively low score, the later or more difficult of the tests of the second group are omitted. The same procedure is followed in the succeeding groups.

The scale is a simple one to administer and to score. The entire directions for giving and scoring, including the table of norms, are included in a small book of fifty-six pages. The scale constitutes an individual test, as in the case of the Binet Scale and the Yerkes Point Scale. It requires less

[1] John P. Herring, *Herring Revision of the Binet-Simon Tests: Examination Manual: Form A.* Yonkers-on-Hudson: World Book Co., 1931.

preparation, however, and, if it proves to be as reliable, it is probably preferable to the more cumbersome and longer individual scales. The intelligence quotient is calculated in the same way as in the case of the Stanford Revision.

3. The United States Army mental tests

At the time that the American Psychological Association, through its president, Dr. Yerkes, and its council, offered its services to the United States Army in the prosecution of the War, and proposed to organize intelligence tests to be given to the army recruits, the chief tests which were in use were the individual age scale and the individual point scale. A considerable number of test groups had been organized, but these were usually not employed extensively except by their originators. A few tests had been administered to groups, but no well-organized group scales had been devised.

The group of army psychologists, who, after a period of experimentation, were entrusted, under the direction of Dr. Yerkes, with the organization of the intelligence examinations, realized that it would be necessary, in order to administer tests on a large scale, to develop a group test. It appears to the lay observer that these psychologists created out of whole cloth radically new methods of examining. On the contrary, they made use of all the earlier experiments with tests which we have reviewed, including the studies of correlation, and simply took the next logical step in advance. This step was taken more quickly than would otherwise have been the case, and the mental-test movement acquired a tremendous impetus as a result of the large number of examinations which were given in the army and of the publicity which it received. One psychologist, Otis, however, was on the point of taking this step himself at the time the army tests were organized, and he contributed his experience and plan to the undertaking of the army psychologists.

The scales which were principally employed in the army were five. Two were scales which we have already described — the Stanford Revision and the Yerkes Point Scale. Two of the others were group tests, and one was an individual performance test.

4. The Army Scale Alpha

The most widely used of the army scales was Scale Alpha. This was a group test which was suitable for administration to men who could understand and could read English. To those who could not understand English because of foreign origin, or because they were illiterate, or because they were mentally defective, was given a second group test which did not involve the use of language. It consisted of a variety of pictures and diagrams. The directions were given by pantomime. The men who failed to make a certain score on this second test, which was called Beta, were given an individual examination. The individual examination was either one of the two which have been mentioned — the Stanford-Binet or the Yerkes Point Scale — or a fifth test which was an individual performance test. These various tests may be described in a little more detail.

On account of the historical importance of Scale Alpha, and because it stands as the type of our group-point scales, we reproduce it almost in full. Revised forms of these tests are still in use and are referred to in the following chapter.

Successive tests are on alternate pages, and tests 5 to 8 are printed upside down in order to prevent the men from looking forward to a new test until the signal is given.[1]

[1] For a brief description of the army tests, and the manual for giving them, see Clarence S. Yoakum and Robert M. Yerkes, *Army Mental Tests.* New York: Henry Holt & Co., 1920. For a full technical account of the derivation of the tests, of the details of the tests themselves, and of the results of the applications in the army, see the official report, Robert M. Yerkes (Editor), *Psychological Examining in the United States Army.* Washington: National Academy of Sciences, Vol. XV, 1921.

It will be seen that the scale consists of eight tests. Test 1 is a so-called *directions test.* Each item is to be marked by the examinee according to directions to be given by the examiner. For example, the directions for the first item of test 1, form 6, are as follows:

"Attention! Attention always means pencils up. Look at the circles at one. When I say 'Go,' but not before, make a cross in the second circle and also a figure one in the third circle. 'Go!'" (Allow not over five seconds.)

The later items of the test are more difficult than the earlier ones. For example the directions for item 12 are as follows:

"Attention! Look at *twelve.* If six is more than four, then, when I say 'Go,' cross out the *five,* unless five is more than seven, in which case draw a line under number *six.* 'Go!'" (Allow not over ten seconds.)

This test is what might be called a test of the mental span, or the ability to keep in mind a number of things at once. It serves also as a means of determining whether the men understand verbal directions, and as a means of weeding out those who do not understand English.

Test 2 is simply a series of arithmetic problems. This might seem at first glance to be merely an educational test. It does, of course, require that the individual shall have had instruction in arithmetic. It was assumed, however, that all of the men being examined had had sufficient instruction to solve these problems if they had the mental capacity to do so. This was probably true for most men, but it was not true for all, particularly some of the foreign-born.

Test 3 is called a test of common sense. It is assumed that every person examined has had the experiences which will enable him to give the correct answer to the questions, provided he has ordinary intelligence.

Form 6 GROUP EXAMINATION ALPHA Score..Rating.

Name..... Date.....

City..... School.....

Age last birthday..... Date of next birthday.....
Month Day

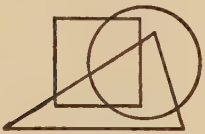
Schooling: Grades, 1. 2. 3. 4. 5. 6. 7. 8: High or Prep. School, Year 1. 2. 3. 4: College, Year 1. 2. 3. 4.

TEST 1

1. ○ ○ ○ ○ ○

2. (1) (2) (3) (4) (5) (6) (7) (8) (9)

3.

4. 

5. ○ ○ ○ *Yes No*

6. ○ ○ ○ ○ ○

7. A B C D E F G H I J K L M N O P

8. ○ ○ ○ *MILITARY GUN CAMP*

9. 34-79-56-87-68-25-82-47-27-31-64-93-71-41-52-99

10.

11. [7F] △ (3) △ 5A (8) [2] △ 6 (9B) [3]

12. 1 2 3 4 5 6 7 8 9

TEST 8

Notice the sample sentence

People hear with the eyes ears nose mouth

The correct word is ears, because it makes the truest sentence.

In each of the sentences below you have four choices for the last word. Only one of them is correct. In each sentence draw a line under the one of these four words which makes the truest sentence. If you cannot be sure, guess. The two samples are already marked as they should be.

SAMPLES { People hear with the eyes ears nose mouth
France is in Europe Asia Africa Australia

1 Boston is in Connecticut Rhode Island Maine Massachusetts 1
2 Euchre is played with dice rackets cards pins .. 2
3 The Arabian is a kind of horse goat cow sheep 3
4 The most prominent industry of Milwaukee is fish brewing flour automobiles 4
5 Turquoise is usually yellow red green blue ... 5

6 The Leghorn is a kind of cow horse fowl granite 6
7 Arthur Brisbane is famous as a newspaper man comic artist athlete actor 7
8 Shoes are made by Swift & Co. Smith & Wesson W. L. Douglas Babbitt Co. 8
9 Blanche Sweet is known as a writer singer suffragist movie actress 9
10 "The makings of a nation" is an advertisement of a tobacco flour beer health food 10

11 Country Gentleman is a kind of wheat corn hay oats 11
12 The artichoke is a vegetable fish lizard snake 12
13 Yale University is at New Haven Annapolis Ithaca Cambridge 13
14 Tokio is a city of India China Egypt Japan 14
15 Diamonds are obtained from mines reefs elephants oysters 15

16 Rodin is famous as a poet painter sculptor composer 16
17 The chameleon is a bird reptile insect fish 17
18 The thyroid is in the shoulder neck head abdomen 18
19 Dioxygen is a disinfectant food product patent medicine tooth paste 19
20 The U.S.S. Michigan is a destroyer monitor submarine battleship 20

TEST 2

Get the answers to these examples as quickly as you can. Use the side of this page to figure on if you need to.

SAMPLES
1. How many are 5 men and 10 men?... Answer (15)
2. If you walk 4 miles an hour for 3 hours, how far do you walk?..... Answer (12)

1. How many are 40 guns and 6 guns?..... Answer (46)
2. If you save $6 a month for 5 months, how much will you save?..... Answer (30)
3. If 32 men are divided into squads of 8, how many squads will there be?..... Answer (4)
4. Mike had 11 cigars. He bought 3 more and then smoked 6. How many cigars did he have left?..... Answer (8)
5. A company advanced 6 miles and retreated 3 miles. How far was it then from its first position?..... Answer (3)

6. How many hours will it take a truck to go 48 miles at the rate of 4 miles an hour?..... Answer (12)
7. How many pencils can you buy for 40 cents at the rate of 2 for 5 cents?..... Answer (16)
8. A regiment marched 40 miles in five days. The first day they marched 9 miles, the second day 6 miles, the third 10 miles, the fourth 9 miles. How many miles did they march the last day?..... Answer (6)
9. If you buy 2 packages of tobacco at 8 cents each and a pipe for 55 cents, how much change should you get from a two-dollar bill?..... Answer (1.29)
10. If it takes 8 men 2 days to dig a 160-foot drain, how many men are needed to dig it in half a day?..... Answer (32)

11. A dealer bought some mules for $900. He sold them for $1,000, making $25 on each mule. How many mules were there?..... Answer ( )
12. A rectangular bin holds 600 cubic feet of lime. If the bin is 10 feet wide and 5 feet deep, how long is it?..... Answer ( )
13. A recruit spent one-eighth of his spare change for post cards and four times as much for a box of letter paper, and then had 60 cents left. How much money did he have at first?..... Answer ( )
14. If 2½ tons of hay cost $20, what will 4½ tons cost?.. Answer ( )
15. A ship has provisions to last her crew of 600 men 6 months. How long would it last 800 men?..... Answer ( )

16. If a train goes 200 yards in 10 seconds, how many feet does it go in a fifth of a second?..... Answer ( )

TEST 7

SAMPLES
sky — blue :: grass — **table** **green** warm big
fish — swims :: man — paper time **walks** girl
day — night :: white — red **black** clear pure

In each of the lines below, the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called.

---

1 dog — bark :: cat — chair **mew** fire house..... 1
2 foot — man :: hoof — corn tree **cow** hoe..... 2
3 dog — puppy :: cat — **kitten** dog tiger horse..... 3
4 wash — face :: sweep — clean broom **floor** straw..... 4
5 door — house :: gate — swing hinges **yard** latch..... 5

6 water — fish :: air — spark man blame **breathe**..... 6
7 white — black :: good — time clothes mother **bad**..... 7
8 boy — man :: lamb — **sheep** dog shepherd wool..... 8
9 roof — house :: hat — button shoe straw **head**..... 9
10 camp — safe :: battle — win dangerous field **fight**..... 10

11 straw — hat :: leather — shoe bark coat **soft**..... 11
12 pan — tin :: table — chair wood legs **dishes**..... 12
13 left — right :: west — south direction east **north**..... 13
14 floor — ceiling :: ground — earth sky hill **grass**..... 14
15 cold — ice :: heat — wet cold steam **stars**..... 15

16 hat — head :: thimble — sew cloth finger **hand**..... 16
17 Monday — Tuesday :: Friday — week Thursday day **Saturday**..... 17
18 lead — bullet :: gold — paper coin silver **copper**..... 18
19 skin — body :: bark — tree dog bite **leaf**..... 19
20 cannon — large :: rifle — ball small bore **shoot**..... 20

21 cellar — attic :: bottom — well tub top **house**..... 21
22 man — arm :: tree — shrub limb flower **bark**..... 22
23 suitcase — clothing :: purse — purchase money string **stolen**..... 23
24 knitting — girls :: carpentry — trade houses boys **lumber** 24
25 arteries — body :: railroads — country train crossing **accident**..... 25

26 ocean — pond :: deep — **sea** well shallow steep..... 26
27 revolver — man :: sting — **gun** hurt bee **hand**..... 27

TEST 3

This is a test of common sense. Below are sixteen questions. Three answers are given to each question. You are to look at the answers carefully; then make a cross in the square before the best answer to each question, as in the sample:

SAMPLE

Why do we use stoves? Because
- ☐ they look well
- ☒ they keep us warm
- ☐ they are black

Here the second answer is the best one and is marked with a cross. Begin with No. 1 and keep on until time is called.

1 If plants are dying for lack of rain, you should
- ☒ water them
- ☐ ask a florist's advice
- ☐ put a fertilizer around them

2 A house is better than a tent, because
- ☐ it costs more
- ☐ it is more comfortable
- ☒ it is made of wood

3 Why does it pay to get a good education? Because
- ☒ it makes a man more useful and happy
- ☐ it makes work for teachers
- ☐ it makes demand for buildings for schools and colleges

4 If the grocer should give you too much money in making change, what is the right thing to do?
- ☐ buy some candy of him with it
- ☐ give it to the first poor man you meet
- ☒ tell him of his mistake

☞ Go to No. 9 above

9 Why are warships painted gray? Because gray paint
- ☐ is cheaper than other colors
- ☐ is more durable than other colors
- ☒ makes the ships harder to see

10 Why should all parents be made to send their children to school? Because
- ☒ it prepared them for adult life
- ☐ it keeps them out of mischief
- ☐ they are too young to work

11 The reason that many birds sing in the spring is
- ☐ to let us know spring is here
- ☒ to attract their mates
- ☐ to exercise their voices

12 Gold is more suitable than iron for making money because
- ☐ gold is pretty
- ☒ iron rusts easily
- ☐ gold is scarcer and more valuable

TEST 6

SAMPLES	2	4	6	8	10	12	*14*	*16*
	9	8	7	6	5	4	*3*	*2*
	2	2	3	3	4	4	*5*	*5*
	1	7	2	7	3	7	*4*	*7*

Look at each row of numbers below, and on the two dotted lines write the two numbers that should come next.

2	3	4	5	6	7	....	....
5	10	15	20	25	30	....	....
10	9	8	7	6	5	....	....
6	9	12	15	18	21	....	....
8	8	6	6	4	4	....	....
3	7	11	15	19	23	....	....
9	1	7	1	5	1	....	....
25	25	21	21	17	17	....	....
4	5	8	9	12	13	....	....
21	18	16	13	11	8	....	....
1	2	4	8	16	32	....	....
3	4	6	9	13	18	....	....
12	14	13	15	14	16	....	....
25	24	22	21	19	18	....	....
16	12	15	11	14	10	....	....
16	8	4	2	1	1/2	....	....
15	16	14	17	13	18	....	....
1	4	9	16	25	36	....	....
21	18	16	15	12	10	....	....
4	8	10	20	22	44	....	....

TEST 4

If the two words of a pair mean the same or nearly the same, draw a line under *same*. If they mean the opposite or nearly the opposite, draw a line under *opposite*. If you cannot be sure guess. The two samples are already marked as they should be.

SAMPLES
good — bad ..... same — opposite
little — small ..... same — opposite

1	cold — hot	same — opposite	1
2	long — short	same — opposite	2
3	bare — naked	same — opposite	3
4	joy — happiness	same — opposite	4
5	find — lose	same — opposite	5
6	shrill — sharp	same — opposite	6
7	minus — plus	same — opposite	7
8	grim — stern	same — opposite	8
9	careless — anxious	same — opposite	9
10	crude — coarse	same — opposite	10
11	commend — approve	same — opposite	11
12	linger — loiter	same — opposite	12
13	agony — bliss	same — opposite	13
14	defective — normal	same — opposite	14
15	competent — qualified	same — opposite	15
16	knave — villain	same — opposite	16
17	null — void	same — opposite	17
18	wax — wane	same — opposite	18
19	adversary — colleague	same — opposite	19
20	altruistic — egotistic	same — opposite	20
21	furtive — sly	same — opposite	21
22	any — none	same — opposite	22
23	asunder — apart	same — opposite	23
24	deplete — exhaust	same — opposite	24
25	superfluous — essential	same — opposite	25
26	recoup — recover	same — opposite	26
27	celibate — married	same — opposite	27
28	recant — disavow	same — opposite	28
29	avarice — cupidity	same — opposite	29
30	aggrandize — belittle	same — opposite	30
31	decadence — decline	same — opposite	31
32	nullify — annul	same — opposite	32

TEST 5

The words A EATS COW GRASS in that order are mixed up and don't make a sentence; but they would make a sentence if put in the right order: A COW EATS GRASS, and this statement is true.

Again, the words HORSES FEATHERS HAVE ALL would make a sentence if put in the order ALL HORSES HAVE FEATHERS, but this statement is false.

Below are twenty-four mixed-up sentences. Some of them are true and some are false. When I say "go," take these sentences one at a time. Think what each would say if the words were straightened out, but don't write them yourself. Then, if what it would say is true, draw a line under the word "true"; if what it would say is false, draw a line under the word "false." If you cannot be sure, guess. The two samples are already marked as they should be. Begin with No. 1 and work right down the page until time is called.

SAMPLES	a eats cow grass	true	false	
	horses feathers have all	true	false	
1	cows milk give	true	false	1
2	write are with to pencils	true	false	2
3	are and apples long thin	true	false	3
4	east the in rises sun the	true	false	4
5	months warmest are summer the	true	false	5
6	wood made carpets are of always	true	false	6
7	known elephant animal an is smallest the	true	false	7
8	water cork on float will not	true	false	8
9	vote children 21 cannot under	true	false	9
10	battleships on seldom sails used are	true	false	10
11	four hundred all pages contain books	true	false	11
12	iron paper made of is filings	true	false	12
13	pays cautious it be to often	true	false	13
14	a general not major a and rank same the of are	true	false	14
15	Washington canal 1776 Panama the in built	true	false	15
16	never deeds rewarded be should good	true	false	16
17	will live bird no forever	true	false	17
18	gases the in Mohawks fighting used poisonous	true	false	18
19	friends in us disaster often false desert	true	false	19
20	external deceptive never appearances are	true	false	20
21	size now of guns use are great in	true	false	21
22	happiness lists great casualty cause	true	false	22
23	always sleeplessness clear causes a conscience	true	false	23
24	inflict men pain needless cruel sometimes	true	false	24

Test 4 is a measure of the ability of the individual to apprehend the relationship of sameness and oppositeness of meaning in words. It is assumed that the persons tested know the meaning of the words. It is obvious, for the latter part of the test at least, that the test is a measure of the understanding of vocabulary as well as a measure of the ability to give opposites.

Test 5 is a measure of the ingenuity of an individual as indicated by his ability to rearrange words and make them into a sentence. To some extent also, of course, it is a measure of information, as in the case of item eighteen.

Test 6 is again a measure of ingenuity, this time in the field of number. It is probably more nearly a pure intelligence test than are some of the others.

Test 7 is a measure of the ability to see relationships. Assuming that the information demanded is at the command of all those who are examined, it has proved to be a good intelligence test. It is called the analogy test, and is the one which was introduced by Yerkes in his point scale in contrast to those which were borrowed from the Binet scale.

Test 8 has been criticized as measuring experience rather than intelligence, but the possession of the information which is demanded by it, assuming the environment of the persons tested has been similar, is regarded as a fair measure of intelligence. However, the test as devised for the Army Alpha discriminates in favor of men as against women.

Each of the tests has a time limit. The various items of test 1, for example, are given from five seconds to twenty-five seconds. The remaining tests have the following time allowances: No. 2, five minutes; No. 3, one and one half minutes; No. 4, one and one half minutes; No. 5, two minutes; No. 6, three minutes; No. 7, three minutes; and No. 8, four minutes. The time limits are so set that but a small percentage, approximately five, shall be able to finish the test. The score which an individual makes, therefore, de-

pends in part upon his speed of performance. However, the tests are not purely speed tests, as is sometimes assumed. First, the tests increase in difficulty, so that if a person's mental capacity is very limited, he begins to slow down sooner than he otherwise would. Second, the rapidity of a person's performance depends in part upon the ease with which he can perform the assigned tasks. The question of power and speed is discussed further in Chapter X.

The examinee is given one point credit for every item which he answers correctly. Since there are 212 items in all of the tests taken together, the highest possible score is 212. Scores of 212 have been reported, but it is very rare that an individual scores above 200.

Detailed directions for scoring the tests are given in the manual. One procedure deserves comment. It will be noticed that in tests 4 and 5 there is an even chance of giving a correct answer if one merely guesses. In each test the examinee is directed, "If you cannot be sure, guess." It is assumed that the examinee will guess on some of the items, and that, upon some of the items on which he guesses, he will obtain a correct answer. This would give him a higher score than he would have if the score is intended to represent only those items to which he knows the answer. A correction is therefore applied to the scores on these tests. The correction assumes that the examinee has made as many correct answers by guessing as he has given wrong answers. His score is, therefore, found by subtracting the number of wrong answers from the number of right answers he has given. We shall examine the validity and usefulness of this procedure in one of the chapters on technique.

There has been a good deal of discussion of the results of the Army Alpha Scale in terms of the letter rating into which the scores were translated. Some have spoken of the letter ratings as though they represented distinct and clearly de-

finable levels of mental capacity. Thus all those who received a grade below C have been spoken of as mentally defective. One writer, on the other hand, has implied that the rating A is a purely arbitrary designation, and explains his contention in this way: The timing of the tests, he writes, was so adjusted that five per cent of the men could finish. To the men who finished were given the grade A. It is therefore due solely to this arbitrary selection of a time limit that approximately five per cent of the men received this grade.[1] As a matter of fact, five per cent of the men did not finish the test as a whole, and the score did not depend merely upon the number of tests which were attempted, but on the number which were correct. The distribution of the scores among the various letter grades was made in a totally different way.

The scores which were assigned to the various letter grades were so adjusted as to give a distribution which approaches the normal distribution more nearly than do the numerical scores. We shall see how this is by examining the scores to which the various letter grades were given, and the distribution of the scores of the men receiving these grades.

TABLE VIII. LETTER RATINGS IN ARMY ALPHA

Letter rating	E and D−	D	C−	C	C+	B	A
Limit of scores	0–14	15–24	25–44	45–74	75–104	105–134	135–212
Range of scores	14	9	19	29	29	29	77
Per cent of principal draft receiving these scores.[1]	7.1	17.0	23.8	25.0	15.2	8.0	4.1

[1] See pages 422 and 800 of the Army Report.

The fact that 4.1 per cent of the men received grade A, therefore, is no mystery. This number received this grade because the range of scores was set at such a point that approximately this number would receive it. It will be noticed that the range of scores for grade A was 77, whereas

[1] Walter Lippmann, "The Mystery of the A Men," *New Republic*, XXXII (1922), 248.

that for grade D was only 9. This is because the scores piled up at the lower end and were much rarer at the higher end. The larger percentage of scores will be seen to be in the divisions at the middle of the scale, and the smaller percentage toward the extremes. If the distribution were entirely normal, there would be an equal percentage in the corresponding divisions ranging from the middle toward the extremes. It is customary to distribute marks or scores in this fashion, and the arrangement of the scores so that they will be so distributed simply means that it is assumed that intellectual capacity occurs among an unselected group of the population in some such form of distribution as this.

There has also been a great deal of confusion concerning the mental ages which were assigned to the various letter grades. The following table gives these corresponding mental ages:

TABLE IX. MENTAL AGES CORRESPONDING TO THE LETTER RATINGS IN ARMY ALPHA

Letter grades	E and D−	D	C−	C	C+	B	A
Corresponding mental ages	0–9.4	9.5–10.9	11–12.9	13–14.9	15–16.4	16.5–17.9	18–

From the table of the distribution of the letter grades of the men in the principal sample given in Table VIII, it will be seen that the sum of the groups below grade C make a total of 47.9 per cent. Nearly 50 per cent of the men, in other words, were rated according to this scheme as below a mental age of thirteen years. A similar method of figuring gives an estimate of 40 per cent as being below a mental age of twelve years. Now it has been the practice of psychologists to interpret a mental age of twelve years, when the individual is mentally mature, as representing marked dullness. Reckoning a normal adult as having a mental age of sixteen, one whose mental age was twelve would have an

I.Q. of 75 ($\frac{12}{16} \times 100$). If we refer back to Terman's distribution of I.Qs., we shall see that only between two and three per cent of children have an I.Q. as low as this.

This enormous discrepancy between the percentage of 40 for adults and two or three for children leads us to inquire how the equivalent mental ages were determined. They were got in this fashion: A carefully selected group of men were given the Army Alpha, and also the Stanford-Binet. The Stanford-Binet mental ages of these men were found. By a comparison of these mental ages with the Alpha scores of the same men, the mental ages which are equivalent to the various Alpha scores were calculated. This procedure assumes that scores made by children and by adults on the same mental test represent equivalent mental capacities. The results of the army test seem to give conclusive evidence that this assumption is not correct. While the discrepancy may be explained in part by other minor factors, the chief explanation must be this lack of equivalence of the results of the test given to children who are in school and are accustomed to doing tasks similar to those demanded by the tests, and to adults who have been out of school for from six to ten years or more, and have lost a good deal of their adeptness for performing tasks which involve clerical skill. It is unsafe, therefore, to interpret the mental-age ratings of adults, when they are obtained in this way, as meaning the same thing as they have been found to mean in our experience with children.

The methods by which the tests were chosen for inclusion in the Alpha scale are instructive from several points of view. In the first place, it should be emphasized that the tests were selected on the basis of careful preliminary trials, and of a statistical tabulation and interpretation of the results of these trials. Each test that was included in the final scale was subjected to careful scrutiny. The correlation

technique which had previously been worked out and applied in the study of single tests was used constantly.

The procedure was to select for preliminary trial a series of tests which had given evidence by previous experimental work of correlating well with general intelligence. These tests were made up into a preliminary scale, called Scale A. They consisted of the following tests: *oral directions*, *memory span*, *disarranged sentences*, *arithmetic problems*, *information*, *opposities*, *practical judgment*, *number completion*, *analogies*, and *number comparison*. Each of these tests was correlated individually with various other measures of intellectual capacity, such as officers' ratings, scores in the Stanford-Binet scale, grade location in the school, and scores in other tests. At the beginning, the plan was to select tests which had high correlation with the outside criteria and low intercorrelation with one another. The reason for the plan to select tests which had a low intercorrelation was largely statistical. Such tests would not be measures of the same thing. A combination of tests with a low intercorrelation, but with a high correlation with the criterion, would from the purely statistical standpoint have a higher correlation with the criterion than a set of tests which measured the same thing and therefore correlated highly with each other. It turned out, however, that the psychological conditions were not in accordance with this statistical demand. The order of the tests as measured by their correlation with the criteria was almost identical with their order as measured by their intercorrelation. This is the same fact as was found by Burt. The detailed evidence of this statement will be presented more fully in one of the chapters on technique. The evidence, then, seems to support the contention of Burt and of Spearman that those tests which are good measures of general capacity measure largely the same factor, or factors, of mental capacity.

5. The Army Scale Beta

In order to provide an examination which could be given to illiterates and non-English-speaking men of foreign birth, the non-language group test Beta was devised. The scale consists of a series of eight tests printed on a paper folder. Each test consists of a series of pictures or drawings which may be understood by a person without the aid of language. The directions are given by means of pantomime. The nature of the scale may be grasped from a brief description of the particular tests.

Test 1 — *Maze test.* This test consists of a series of lines which form five mazes. In each case the examinee is required to draw a line by the shortest route from the left-hand side to the right-hand side of the maze, without going into any blind alleys. This test was suggested by an earlier one devised by Porteus.[1]

Test 2 — *Cube Analysis.* This test consists of a series of drawings, each one of which represents a series of cubes piled upon one another in regular fashion. Some of the cubes are hidden from view and the examinee is required to tell how many cubes are in the pile.

Test 3 — *XO Series.* This test consists of a series of arrangements of the letters X and O. At the end of each series are a number of blanks which are to be filled out according to the same arrangement.

Test 4 — *Digit Symbol.* A substitution test similar to that used by Healy, Pyle, and others.

Test 5 — *Number Checking.* This test consists of a series of pairs of numbers, beginning with short ones and ending with long ones. The subject is required to check those pairs which are alike.

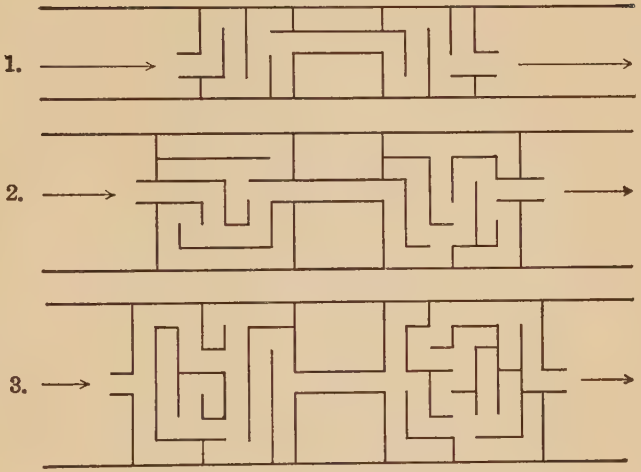
Test 6 — *Pictorial Completion.* A series of pictures, each with one part left out, which is to be supplied by the examinee.

Test 7 — *Geometrical Construction.* This is derived from the form-board test. It consists of a number of items. Each item contains a square and a number of figures which, when put together in the proper way, compose the square. The subject is to draw a line in the square to indicate how the figures might be arranged in it.

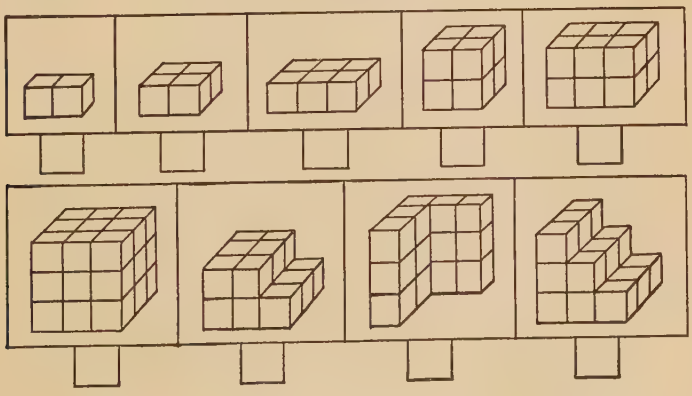
[1] S. D. Porteus, "Mental Tests for Feeble-Minded: A New Series," *Journal of Psycho-Asthenics,* XIX (1915), 200–13.

GROUP EXAMINATION BETA

TEST 1



TEST 2



TEST 3

1.

2.

3.

4.

5.

6.

TEST 4

1	2	3	4	5	6	7	8	9
—	И	Ǝ	L	U	O	Λ	X	=

1.

3	1	2	1	3	2	1	4	2	3	5	2	9	1	4

2.

6	3	1	5	4	2	7	6	3	8	7	2	9	5	4

3.

6	3	7	2	8	1	9	5	8	4	7	3	6	9	5

TEST 5

650	 650	10243586	 10243586
041	 044	659012534	 659021354
2579	 2579	388172902	 381872902
3281	 3281	631027594	 631027594
55190	 55102	2499901354	 2499901534
39190	 39190	2261059310	 2261659310
658049	 650849	2911038227	 2911038227
3295017	 3290517	313377752	 313377752
63015991	 63019991	1012938567	 1012938567
39007106	 39007106	7166220988	 7162220988
69931087	 69931087	3177628449	 3177682449
251004818	 251004418	468672663	 468672663
299056013	 299056013	9104529003	 9194529003
36015992	 360155992	3484657120	 3484657210
3910066482	 391006482	8588172556	 8581722556
8510273301	 8510273301	3120166671	 3120166671
263136996	 263136996	7611348879	 76111345879
451152903	 451152903	26557239164	 26557239164
3259016275	 3295016725	8819002341	 8819002341
582039144	 582039144	6571018034	 6571018034
61558529	 61588529	38779762514	 38779765214
211915883	 219915883	39008126557	 39008126657
670413822	 670143822	75658100398	 75658100398

TEST 6

1.

2.

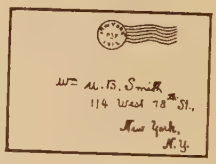
3.

4.

5.

6.

7.

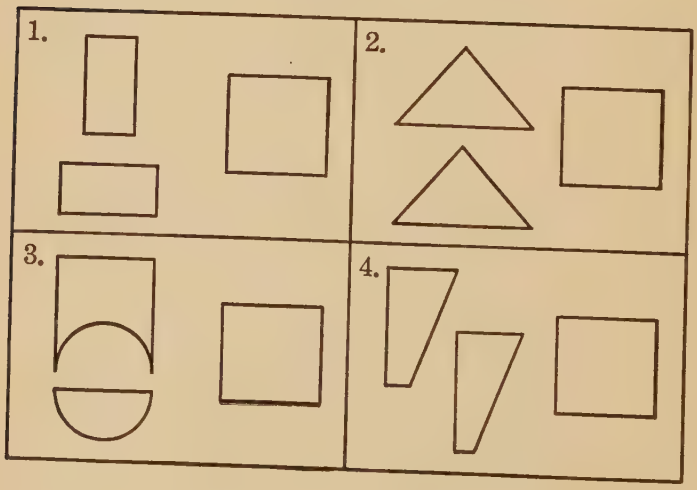
8. 

9.

10.

11.

TEST 7



Since the Beta scale is the prototype of the group non-language scales, as Alpha is the prototype of the group language scales, it is reproduced in full.

This scale was found to be reasonably satisfactory, although it did not give as accurate measurements as the Alpha scale. It was difficult to give the directions by pantomime, and variations in procedure were likely to occur. It has stimulated the development of a considerable number of similar scales for application to school children, particularly those for the primary grades in which children cannot read readily. A modern revision of the test has been made to be administered with oral directions.

6. The performance scale examination

In case the recruit had made a low score on the Alpha scale and the Beta scale, he was given one of the three individual examinations. In addition to the Stanford Revision and the Yerkes Point Scale, a performance scale was devised. This consists of tests which require the individual to react to problems which are presented, not in the form of words, but of concrete objects. In some cases these objects are drawings, and in other cases they are composed of solid objects. The nature of the scale may be briefly described.

Test 1 — *The Ship Test* (Knox). This consists of a rectangular picture pasted on a thin board, and cut up into ten pieces. The pieces are to be arranged by the subject so as to make the picture. (See p. 180 for illustration.)

Test 2 — *Manikin* (Pintner) *and Feature Profile* (Knox). These tests were derived from the series by Knox and Pintner referred to in Chapter VII. They are simple construction puzzles, one representing a face and the other a man.

Test 3 — *Cube Imitation* (Knox). This is the Knox test described in Chapter VII.

Test 4 — *Cube Construction* (Goddard). This test requires that

the subject shall put together small cubes painted on certain surfaces in such a way as to make a larger block painted on certain of its surfaces.

Test 5 — *Form Board* (Dearborn). This is somewhat similar in its make-up to construction puzzle B of the Healy-Fernald series. Both the original and the revised form used in the army were devised by Dearborn and his associates.[1]

Test 6 — *Designs* (Terman). A series of figures are shown to the subject which he is to copy from memory as nearly as possible.

Test 7 — *The Digit-Symbol Test.* The same test as was used in Beta.

Test 8 — *The Maze* (Porteus). These mazes are similar in principle to the ones used in Beta.

Test 9 — *Picture Arrangement* (Bowler, Whipple). A series of "Foxy Grandpa" pictures placed out of order. They are to be placed in order so as to make the sequence.

Test 10. — *Picture Completion* (Healy). Similar to the last picture completion test of the Healy series.

A third individual test, which was given in a few special instances to test mechanical ability, was the Stenquist Mechanical Aptitude Test. This test is described in the chapter on tests for the analysis of mental capacity.

7. *The uses of mental tests in the army*

The psychological committee planned to use mental tests primarily to detect drafted men who were too low-grade mentally to make satisfactory privates, to discover those who were mentally unstable and might prove incorrigible, and if possible to select exceptional men who might be used for tasks demanding a high degree of intelligence. The uses to which the tests were actually put are classified briefly by Yoakum and Yerkes, as follows:[2]

[1] W. F. Dearborn, J. E. Anderson, and A. O. Christiansen, "Form Board and Construction Tests of Mental Ability," *Journal of Educational Psychology*, VII (October, 1916), 445–58.

[2] Yoakum and Yerkes, *op. cit.*, pp. xii and xiii.

1. The assignment of an intelligence rating to every soldier on the basis of systematic examination.
2. The designation and selection of men whose superior intelligence indicates the desirability of advancement or special assignment.
3. The prompt selection and recommendation for development battalions of men who are so inferior intellectually as to be unsuited for regular military training.
4. The provision of measurement of mental ability which enabled officers to build organizations of uniform mental strength or in accordance with definite specifications concerning intelligence requirements.
5. The selection of men for various types of military duty or for special assignment, as for example, the military training schools, colleges or technical schools.
6. The provision of data for the formation of special training groups within the regiment or battery, in order that each man may receive instructions suited to his ability to learn.
7. The early discovery and recommendation for elimination of men whose intelligence is so inferior that they cannot be used to advantage in any line of military service.

The use of the tests as one of the means of selection of officers is based upon the superiority of officers in the test ratings. The distribution of the scores of different groups of men is shown in Fig. 2.

As the result of the experiments with the tests the following summary is given: During a specimen six months' period, one half of one per cent were reported for discharge because of mental inferiority, six tenths of one per cent were recommended for assignment to labor battalions because of low-grade intelligence, and six tenths of one per cent were recommended for assignment to development battalions, in order that they might be more carefully observed and given preliminary training. The purpose of this training was to discover means of giving the men training which would fit them to be useful soldiers. The army psychologists believed that there were nearly three per cent of the men who were

so low-grade mentally that they were not of sufficient service to compensate the Government for the expense necessary to equip and train them for service. Among the directions

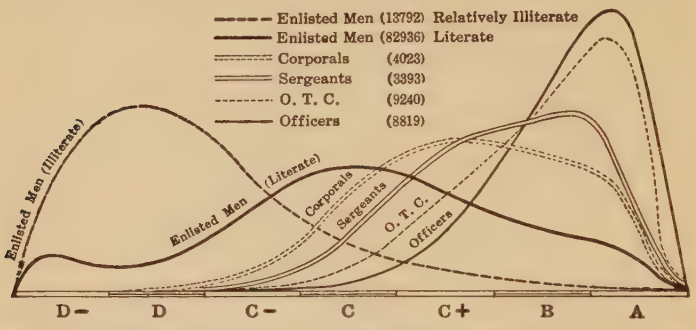


FIG. 2. THE DISTRIBUTION OF INTELLIGENCE RATINGS IN TYPICAL ARMY GROUPS, SHOWING THE VALUE OF THE TESTS IN THE IDENTIFICATION OF OFFICER MATERIAL

From Yoakum and Yerkes, *Army Mental Tests*. Henry Holt & Co., 1920. By permission of the publishers.

which were issued by the psychological service for the use of results of the psychological examinations, the following will throw additional light upon the application of these tests in the army:

First, the tests were not designed to be a substitute for other methods of judging a man's value to the service. They were not intended to measure character traits, such as "loyalty, bravery, power to command, or the emotional traits that make a man carry on." Intelligence, however, was regarded as the most important single factor in efficiency. Second, it was expected that commissioned officers would be found chiefly among the men who received the grades of A or B. Men with grades below C+ were expected rarely to have the capacity for success in officers' training schools. Non-commissioned officers, furthermore, were expected to be chosen chiefly from the men whose grades were C+ or higher.

In selecting men for positions of special responsibility which corresponded to particular occupations of civil life, those men were first to be selected whose intelligence rating was above the average of men in that occupation. The intelligence ratings of the men in the army were classified according to their civil occupations.

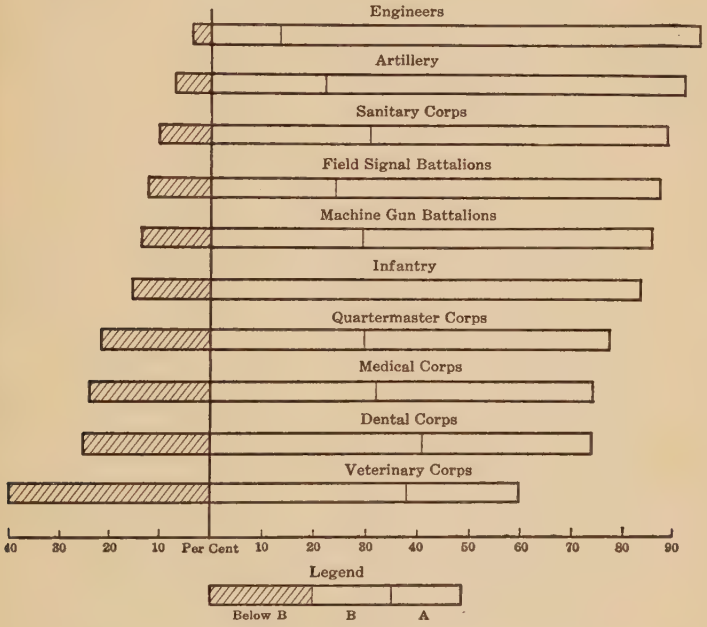


FIG. 3. INTELLIGENCE OF OFFICERS IN DIFFERENT ARMS OF THE MILITARY SERVICE

From Yoakum and Yerkes, *Army Mental Tests*. Henry Holt & Co., 1920. By permission of the publishers.

It was directed that men be assigned to permanent organizations with a view to making these organizations equal in average intelligence. The only exception to this was the case of certain arms of the service which were found to re-

quire greater than the average intelligence. Such arms were the signal corps, machine guns, field artillery, and engineers. The variation in the rating of men in different arms of the service is shown in Fig. 3.

It is evident that the ability which is measured by the tests existed in different degrees in officers and men, and in the men of different branches of the service. The tests constituted, therefore, one of the means by which the men who were fitted for the successful performance of different functions in the army might be selected.

Chapter VI

SURVEY OF POINT SCALES

In this chapter we shall discuss, first, the main facts concerning the later development of the group point scales which are now available for use in the schools. We shall next consider the criteria which should be kept in mind in the choice of a scale to be used in the school. Finally, we shall present a selected list of the chief existing group point scales.

1. Later development of group tests

The army testing work bore fruit very rapidly in group point scales for use in schools and colleges. The War had not closed when Otis published his advanced examination. He had been working upon this test before the War opened, and published it in May to June, 1918. Within five years there had appeared approximately fifty such scales for schools and colleges, and the number has steadily increased down to the present. The Otis scale contains ten tests. It requires a full hour to give and is designed for the high school. It has had rather wide application, but is being displaced by other tests which do not require so much time and can more easily be given, among them Otis's own Higher Examination.

Beginning in 1917, Whipple began an elaborate study of the value of many of the various single tests which had been developed up to that time, for the purpose of selecting children for a special class for gifted pupils. At the completion of this study, which was reported in 1919, in his Classes for Gifted Children, he organized the tests which he found to be most suitable into a series of group tests. These constitute his Group Tests for Grammar Grades.

At about the same time as the Otis test the Group Point Scale for Measuring General Intelligence was published by S. L. and L. W. Pressey. The Pressey test was also designed for use in the high school and has had rather wide application, particularly in surveys of the secondary schools of Indiana. It has sometimes been used for testing applicants for college entrance.

In 1919 and 1920, respectively, there appeared, as a direct outgrowth of the army tests, the Haggerty Delta 1 and Delta 2, and the National Intelligence Test. Haggerty had been engaged with the psychological service in the army, although not in the psychological testing. He worked out the two above mentioned scales in connection with the Virginia School Survey. Delta 1, one exercise of which is reproduced in Fig. 4, recalls the Army Beta Test, which will be remembered as a non language test. Delta 2 recalls Army Alpha. These scales were carefully adapted to the mental development of children in the primary grades and in the upper grades respectively.

The National Intelligence Tests were worked out by a committee consisting of Haggerty, Terman, Thorndike, Whipple, and Yerkes. This committee was granted the sum of $25,000 by the General Education Board to conduct researches and to devise tests which should be more highly refined than was possible without such extensive investigation. There are two scales, Scale A and Scale B, and two parallel forms of each scale. Other parallel forms are in process of development. The constituent tests are for the most part similar in character to those in Army Alpha. A feature which distinguishes these scales from most others is that each test is preceded by a practice exercise. Partly on account of the prestige of the committee which organized the tests, they have had very wide application, and the norms

FIG. 4. ILLUSTRATION OF PART OF A NON-LANGUAGE TEST FOR PRIMARY GRADES

From The Haggerty Intelligence Examination, Delta 1. (Reproduced with the permission of the author.)

which are furnished for them are based upon large numbers of children — about four thousand for each grade or age. It is not known, however, whether the test is more valid or more reliable than other similar scales.

A group of scales may be mentioned because they possess the common characteristic of being in nonlanguage form. The Myers Mental Measure is a brief test covering four pages, and designed to furnish intelligence ratings on individuals all the way from kindergarten to the graduate school. It is therefore very steeply graded; each of the four tests contains some very easy and some very difficult items. The Pintner Non-Language Test is suited to the upper grades of the elementary school and is designed particularly for the purpose of testing children who are not familiar with the English language, either because of foreign extraction, or because of deafness or some other disability. It is also designed to supplement the language scales. The Thorndike Non-Language Test was prepared for adults and aims to provide a score which shall be based not on language ability but on ability to deal with problems presented in more concrete form. This test is now out of print.

A rather widely used maze test, devised by Porteus,[1] is thought by its author to measure foresight or circumspection. The mazes are reproduced in Fig. 5. The subject is directed to trace the shortest line from the entrance, at S, to the other opening. This test correlates rather closely with an intelligence test, and the opinion that it measures a special capacity is based on introspection and analysis rather than on statistical evidence.

A test which is sometimes included among perception tests but which should probably be regarded as measuring higher ability is the Kohs [2] test. This test, which has been

[1] S. D. Porteus, "Motor Intellectual Tests for Mental Defectives," *Journal of Experimental Pedagogy*, III (1915), 127–35; "The Measurement of Intelligence: Six Hundred and Fifty-Three Children Examined by the Binet and Porteus Tests," *Journal of Educational Psychology*, IX (January, 1918), 13–31.

[2] S. C. Kohs, *Intelligence Measurement.* New York: Macmillan Co., 1927.

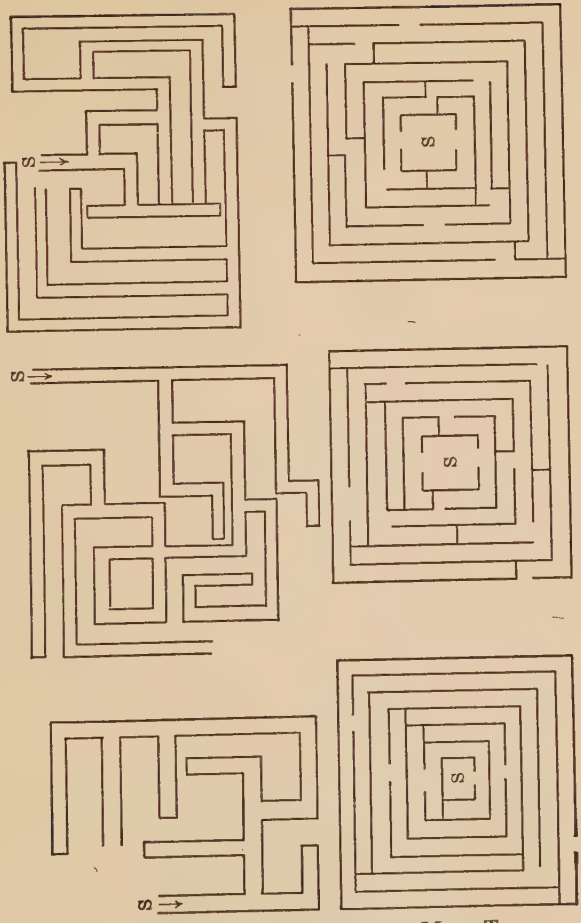


FIG. 5. SPECIMENS OF THE PORTEUS MAZE TEST

(Copied by permission of the C. H. Stoelting Co.)

elaborately worked out, requires the child to copy a pattern by putting together blocks of various colors. Kohs believes that it measures the higher intellectual processes, and

defines these processes as consisting of analysis and synthesis. He finds that his test correlates well with general tests of intelligence. This correlation, of course, raises the question whether the test is really a specialized or a general test.

A nonlanguage test designed to measure motor ability as well as two abilities which are commonly included in intelligence, namely, "feature discrimination" and maze solving, is the Michigan Non-Verbal Series, by Edward B. Greene. The feature discrimination test is somewhat like a test devised by Stuart C. Dodd for an "International Test," which was published only in an experimental edition. It contains groups of diagrammatic sketches of human faces. The testee is to draw a circle around the feature of one face which makes it different from the others. The maze test is similar in principle to that used in the Army Beta.

A novel nonlanguage test, adapted especially to younger children, is a drawing test devised by Goodenough. The child is merely asked to draw a human figure, and the drawing is scored according to a standardized scale of items.

The most complete nonlanguage group test adapted to a wide range of ages is the Chicago Non-Verbal Examination, by Andrew W. Brown, published in 1936. This is a collection of ten tests using familiar types of material but with new content. For example, the first test requires the writing of digits under appropriate symbols according to a key. The second is a classification test in which the testee marks one of a group of drawings which differs from the rest. The third requires the calculation of the number of blocks in a pile represented by a two-dimensional drawing. The seventh requires the testee to indicate the proper sequence of a group of four or five drawings, and so on. This test will doubtless prove useful for testing persons in homes where a foreign language is spoken, for the deaf, and for those who are deficient in the ability to read.

A group of nonlanguage tests which do not use pencil and paper but require rather the manipulation of actual objects are designated performance tests. A number of performance tests have already been mentioned: the Army Performance Scale, the Pintner-Paterson Scale, the Knox tests used at Ellis Island, the form boards and the Kohs Block Design Test.

Three other performance scales may be mentioned. The Merrill-Palmer Scale, arranged by Rachel Stutsman, is especially adapted for use in the pre-school period. It includes, besides the performance tests, a few language tests. The performance tests include three groups, the all-or-none tests, the form boards and picture tests, and other tests of motor co-ordination. In the first group the child is directed to perform certain simple acts, such as throwing a ball, cutting with scissors, opposing thumb and fingers. In the second, he is required to respond to a selected list of form boards, chosen from previous test groups. In the third, he is given a series of tasks, such as fitting cubes in a box. The scale requires a rather elaborate set of materials. The correlation of the scores with language scales such as the Binet is reported by the author to be about .80, but a lower figure is reported in a study by Kawin.[1] The cause of the discrepancy is not certain.

The Arthur Point Scale of Performance Tests is designed for clinical use over a wider range of ages, with norms from six to twenty-one. The scale is composed of a number of form board and performance tests designed by previous investigators, including such tests as the Knox Cube, Seguin Form Board, Healy Picture Completion, I and II, Porteus Maze, Kohs Block Design, and others. There are two forms of equivalent value.

[1] Ethel Kawin, *Children of Preschool Age.* Chicago: University of Chicago Press, 1934.

Another performance scale made up in much the same way and including some of the same tests is the Cornell-Coxe Performance Ability Scale. In addition to five strictly performance tests it includes the digit-symbol test from the Army Beta and the test of memory for designs.

A new development since 1925 is a series of developmental schedules for infants and pre-school children. The exact measurement of development by comparison of the infant's behavior with an inventory of activities and the average dates of their appearance was inaugurated by Gesell, who published a Pre-School Child Development Scale in 1925.[1] This scale consists of an inventory of the activities of the infant with the schedule of the average ages at which they appear. Some of the activities are spontaneous or appear in response to the general situation; others are elicited by the presentation of an object by the examiner. Such presentation constitutes a test. The inventory of Gesell is the pioneer in the field of infant testing and has been freely drawn upon by other workers. These workers have established scales by describing scoring procedures and methods of getting composite scores.

A number of these infant scales may be mentioned. One by Linfert and Hierholzer,[2] covering the first year of life, was published in 1928. This scale contains two series, one for the first half-year and the other for the second half-year. A more recent test for the first year has been produced by

[1] Arnold Gesell, *The Mental Growth of the Pre-School Child.* New York: Macmillan Co., 1925. The description of behavior at successive age periods within the first year is amplified in the following: Arnold Gesell and Helen Thompson, *Infant Behavior: Its Genesis and Growth.* New York: McGraw-Hill Book Co., 1934; and Arnold Gesell and Helen Thompson, *The Psychology of Early Growth: Including Norms of Infant Behavior and a Method of Genetic Analysis.* New York: Macmillan Co., 1938.

[2] Harriette-Elise Linfert and Helen M. Hierholzer, *A Scale for Measuring the Mental Development of Infants During the First Year.* Baltimore: Warwick & York, Inc., 1928.

Nancy Bayley. This scale, the California First-Year Mental Scale, has been very carefully standardized.[1] It consists of 115 items, listed in the order in which they are passed and each given an absolute scale value. The scoring is in terms of the sigma score, which is the ratio of the differences between the child's score and the mean for his age to the standard deviation of scores for his age. A third scale is called the Iowa Tests for Young Children.[2] It extends through the first two years. A test for still older children, from one and one-half to six years, is the Minnesota Preschool Scale.[3] This is of the Binet type, but scored in C-points rather than by direct calculation of mental ages. M.A. and I.Q. equivalents can be obtained from conversion tables. There are two forms. The test was standardized on one hundred children of each half-year of age, selected from families representing the same occupational sampling as in the population at large.

The primary tests, like the first one which was devised by Haggerty, are of the nonlanguage type. A nonlanguage test is, of course, necessary since children in the primary grades either cannot read or read so haltingly that a printed group test in language form would be almost entirely a test of reading ability. A number of excellent tests for the primary grades have been published, built on the same general model as the Haggerty tests. They are too numerous to be described *in toto*. A selected list is given at the end of the chapter. A few may be mentioned because of their distinctive features. The Kingsbury Primary Test

[1] Nancy Bayley, *The California First-Year Mental Scale.* Berkeley: University of California Press, 1933.

[2] Eva A. Fillmore, *Iowa Tests for Young Children.* University of Iowa Studies in Child Welfare, Vol. XI, No. 4. Iowa City: University of Iowa, 1936.

[3] Florence L. Goodenough, Josephine C. Foster, and M. J. Van Wagenen, *Minnesota Preschool Scale.* Minneapolis: Educational Test Bureau, 1932.

contains four parts. The first is a more or less conventional directions test. The other three were consciously designed to represent, in the form of pictures, the mental processes which have been found to give good results when given in language form in general intelligence tests. The test which is illustrated in Fig. 6 is a completion test, made after the analogy of the language completion test. Each figure contains a series of drawings, with blank spaces which are to be filled in by the child. For example, one of the easiest contains a series of circles of progressive size; the last space is left vacant and in it the child is to draw a circle larger than the one preceding. The Detroit Kindergarten Test is especially noted because it is adapted to the kindergarten level. The Cole-Vincent Test is designed for children entering school. A later test also designed for children entering school is the Metropolitan Readiness Test.[1] While this is not called an intelligence test it does not differ in any essential respect from tests which are so called. The Kuhlmann-Anderson test is distinctive in that it extends from the first grade to the adult level. It includes a series of tests of increasing difficulty, grouped in booklets which overlap and are suited to successive groups of ages. The Otis Quick-Scoring Mental Ability Test, Alpha Test, is furnished with a stencil which facilitates scoring.

Among the tests for intermediate and upper grades the National Intelligence Test is declining in use because of its difficulty of administration and the appearance of more convenient tests. Among the less expensive and more easily administered tests for this level are the Henmon-Nelson Test of Mental Ability and the Otis Quick-Scoring Mental Ability Test, Beta Test. The Henmon-Nelson Test uses a quick-scoring device called the Clapp-Young Self-

[1] Gertrude H. Hildreth and Nellie L. Griffiths, *Metropolitan Readiness Tests.* Yonkers-on-Hudson, New York: World Book Co., 1933.

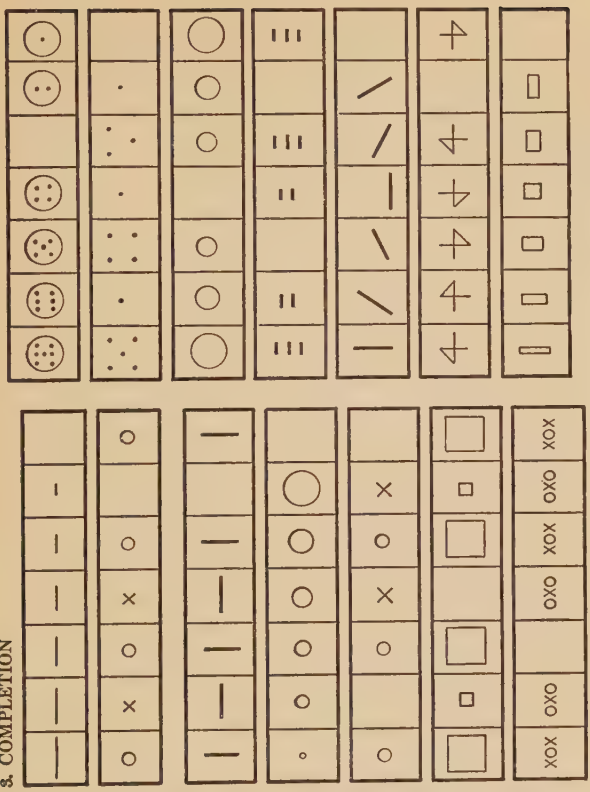


FIG. 6. ILLUSTRATION OF ONE EXERCISE OF THE KINGSBURY PRIMARY TEST
Designed especially to duplicate pictorially the situations set up in the most successful language tests.
(Reproduced with the permission of the author.)

Marking Device, which is described on page 160. The McCall Multi-mental Scale falls in this category, but it is confined to a single process, that of classification of words.

Among the more widely used tests at the high-school level are the Terman Group Test of Mental Ability and the Otis Self-Administering Test of Mental Ability, Higher Examination. Widely used tests generally available at the college level are the Army Alpha (one of the revised forms)

and the Psychological Examination for High School Graduates and College Freshmen, devised by Thurstone and published by the American Council on Education. New tests which give promise of being widely used at the high-school and college level are the Henmon-Nelson tests and the Otis Quick-Scoring Test, Gamma Test.

The earlier tests follow the organization of the army group test, that is, the scale is made up of a series of graded tests and all the similar items are segregated. For example, the arithmetic tests are all together, and so on. Furthermore, separate directions are given for each of the individual tests of the scale and each one is timed separately. In 1919 Thurstone put out a test for high-school graduates and college students which was designed on a different plan. This scale was illustrated in Chapter I. It contains a variety of tests, not segregated, but arranged in rotation, or in cycles. Easy examples of the various tests are placed at the beginning of the series, then slightly harder examples, and so on. When a new variety of test is introduced, it is explained by an example. Thus the test is self-explanatory. The examiner gives the directions once for all at the beginning and keeps time for the entire scale as a unit instead of timing the various parts separately. The Army Alpha Scale has been arranged in this spiral fashion and is called *Scrambled Alpha*. Otis, Henmon-Nelson, and others have more recently adopted this form of organization.

A number of scales have been devised in which an effort is made to measure the relationship between capacity and performance. The first scale in which this measurement was attempted was the Illinois Examination, which was published in 1920. This examination consists of two parts. One part is made up as an ordinary intelligence examination, and the other part consists of an examination of two of the school subjects, arithmetic and reading. In scoring this

test the two parts are kept distinct. The intelligence score is found in terms of mental age, and the achievement score in terms of achievement age. Mental age means the same thing as in any intelligence test. Achievement age is found by comparing the score which the pupil makes with a series of achievement norms. These achievement norms are the median scores made by the pupils of the various mental ages. (The usual practice is to use chronological ages.) After the mental age and the achievement age are found, the next step is to divide the achievement age by his mental age. This gives the achievement quotient. The achievement quotient then is the ratio between what the pupil accomplishes and what it is assumed he could accomplish because of his intelligence rating. Other tests which are divided into two parts, the one an intelligence test and the other a subject-matter test, are the Mental-Educational Survey Test, by Pintner, and the New Jersey Composite Test. Another test of this character is the Otis Classification Test. The test is made up of two parts, the first part an Omnibus Achievement Test, and second an Omnibus Mental Test. The arrangement is the same as in the author's self-administering test.

A new scale is in process of standardization by M. A. Wenger and the writer which contains features never before combined into one scale. The original purpose of the design of the scale was to produce a test for a long-time longitudinal growth study which would be an improvement on the battery used by Flory and the writer.[1] This battery consisted of the VACO tests (vocabulary, analogies, completion, and opposites). These tests were given indi-

[1] Frank N. Freeman and Charles D. Flory, *Growth in Intellectual Ability as Measured by Repeated Tests.* Monographs of the Society for Research in Child Development, Vol. II, No. 2. Washington: National Research Council, 1937.

vidually and did not extend below age five. Moreover, they were all language tests. The new scale contains ten tests, four nonlanguage and six language. They meet the following criteria: (a) they are generally recognized as suitable for measurement of intellectual behavior, (b) they have a high positive correlation with the general factor in the Holzinger bi-factor analysis of the preliminary data, (c) they are applicable over a wide age range (three or four to adult) but demand presumably the same or similar mental processes, (d) they are adapted to multiple response scoring, (e) they have intrinsic interest to the testee, (f) they permit of group presentation without time limits (except that individual presentation is required at the lower ages), and (g) they are arranged for electrical machine scoring.

The list of the tests with a brief description of each is as follows:

Nonlanguage Tests

1. *Picture sequence.* Drawings from newspaper cartoons depicting the dog, Napoleon, and his master, Uncle Elby, have been disarranged and presented in sets of three, four, and five. The task is to determine and underscore the second picture of the true sequence.[1]
2. *Form board.* Four figures, a square, a triangle, a hexagon, and a trapezoid, have been cut into variously shaped parts. Each test item consists of the parts for one figure, which the testee is required to identify by underscoring one of four miniature replicas.
3. *Series correction.* A modification of the Army X–0 series in which the task is to discover and underscore the one of fifteen capital letters which destroys the series. At the individual presentation level pictures of animals, birds, flowers, and toys are substituted for letters, and the beginning series contain only seven units.

[1] The authors are indebted to Mr. Clifford McBride, the creator of Napoleon and Uncle Elby, and to Mr. Arthur J. Lafave, of the Lafave Syndicate, for their very generous permission to use their material.

4. *Pattern analogies.* The majority of items for this test have been taken without change from the American Council on Education Psychological Examinations, prepared by Thurstone and Thurstone.[1] Some new items have been added.

Language Tests

5. *Opposites.* Following Freeman and Rugg [2] an attempt has been made to present five alternate responses to the stimulus words, each of which is, in a sense, an opposite, but one of which is the direct opposite. At the individual presentation level the stimulus word is given orally, and a verbal response is required.
6. *Word grouping.* The familiar word classification test in which each item consists of five words, four of which have some common characteristic. At the individual presentation level line drawings of objects replace words.
7. *Problems.* This test takes its character from the graded series of questions presented by Burt [3] in his test of reasoning ability and is supplemented by syllogistic items and by others similar to those contained in the Army Alpha test of judgment. Four possible responses are provided for each item. At the individual presentation level the problems are presented verbally with the aid of pictures.
8. *Analogies.* Items for this test have been selected from those standardized by Pintner and Renshaw and have been supplemented by new items at the individual presentation level which are presented in picture form.
9. *Sentence completion.* Consists of modified items from the I.E.R. Inventory, standardized by Thorndike for his CAVD scale. At the levels for group presentation multiple responses have been provided. The testee is required to underscore one of five letters or groups of letters which constitutes the initial letter of a word or words adequate to complete the sentence.

[1] The authors are indebted to Drs. L. L. Thurstone and Thelma Gwinn Thurstone and to the American Council on Education for their permission to use their test items.

[2] From an intelligence test now out of print.

[3] The authors are indebted to Professor Cyril Burt for his permission to make use of his test.

10. *Vocabulary.* Since the I.E.R. vocabulary items already are in multiple choice form no modification was necessary for their inclusion in the present battery. Items have been selected in terms of the difficulty values given by Thorndike.[1]

While it is, strictly speaking, only an achievement test and therefore does not belong in the discussion of mental tests, the Stanford Achievement Test may be mentioned here because it furnishes a composite measure of achievement and may be used in combination with a mental test in a similar fashion to the Illinois Examination or the Pintner or Otis tests.

This brief survey has not been at all exhaustive, and many tests have not been mentioned which are perhaps as serviceable as those which have been singled out for special discussion. A selection of the tests which have been mentioned is based largely either upon their historical importance or upon the fact that they contain unique or uncommon features.

We may turn next to a consideration of those characteristics which are important in determining the selection of a test.

2. Criteria for the choice of a mental test

Price. The price of mental tests varies greatly from one cent apiece at one extreme to one dollar at the other. Since the price of the test may be a determining factor in the decision whether a testing program may be launched upon or not, this is an important consideration. Furthermore, the value of the service of tests for a particular purpose does not always vary directly with price. The price range of the group tests at the present time is from four to six cents apiece. The price of a test should be considered in

[1] Items from the I.E.R. sentence completion and vocabulary tests are used with the kind permission of Professor Edward L. Thorndike.

one of two circumstances: First, in case two or more tests are equal in other respects but differ in price; or second, in case there is a definite limitation upon the appropriation for the testing program. However, in calculating the cost of the entire administration of the test, other items must be considered, such as the time required to administer or to score the test. These will be mentioned below.

Completeness and convenience of material, and fullness, simplicity, and clearness of directions. There has been such keen competition in recent years in the production of standardized tests that the material and directions for the tests, including the materials necessary to score them easily and quickly, have come to be pretty thoroughly standardized. Unless there is some very strong reason to the contrary, it is by all means advisable to select a test for which the materials and directions have thus been worked out.

Adaptation to ages or grades to be tested. In the list of tests on pages 164–68 they are classified roughly according to the periods of school in which they are designed to be used. The tests, in general, may be classified as belonging to one of five periods: pre-school, primary grade, upper grade, high school, and college. In some cases the pre-school and the primary period may be served by the same test, and in some cases the same test may be used in the high school and in the college. A few tests have been devised which are designed to cover a wider range in ages than is common. Such tests must either be made longer than the ordinary test, or they must have fewer items which are suited to the stage of development at a particular age. In the first case the test would be less reliable because based upon fewer numbers of items. In general it is probably desirable to use tests which are adapted to a fairly narrow range.

Appeal to the child. The typical modern intelligence test is very interesting to the child. Unless there is some special condition which makes the child nervous, he enjoys taking a test, particularly if he is at all accustomed to it. It is not at all difficult, therefore, to find tests which will be entirely satisfactory from this point of view for any stage of development.

The content of the test. This refers to the subject-matter of the tests, or perhaps to the mental processes which they are supposed to measure. This feature is not a very important practical consideration in the choice of tests. As we shall see in the discussion of the content of the tests, in the chapters on technique, various tests have been shown to be of about equal value as constituents of an intelligence scale. Furthermore, most of the scales use, at least in part, the same tests. We cannot, as it has sometimes been thought, determine just what specific mental capacity is measured by a particular test. In fact, it is probable that each of the tests measures a variety of mental capacities.

The length of the scale. The length of the scales varies considerably, from those which contain only four or five tests, and which can be given in fifteen or twenty minutes, to those which require from an hour to three hours to give. Theoretically, up to a certain point the increase in the length of a test adds to its reliability and therefore to its validity. This is because chance errors are diminished by an increase in the number of responses which the child makes. For example, if the test calls for the possession of a number of items of information, there is a chance that a given individual might possess or fail to possess a given item accidentally. The crucial question is, What is the point at which we begin to have rapidly diminishing returns from an increase in the length of the test? Apart from the theoretical question, it may be necessary in some cases to select a short test for

practical reasons. Again, the length of a test should be measured not so much by the total amount of time required to give the test as by the amount of time the pupil actually spends in working upon it. A test which is organized on the omnibus or spiral plan is more economical of time than one which is made up of a series of segregated tests, each one of which has to be presented to the class individually. So far as length is a factor in reliability, the reliability of the test depends not upon the amount of total time required to give it but upon the amount of time the pupil actually spends upon it and the number of items of which it is composed.

Ease and simplicity of administration. This refers to the ease of preparing to give the test and of presenting it to the class. Tests differ in this respect, although most of the newer tests are very easy to prepare and to administer. In general, the omnibus or self-administering tests are much simpler to give than the tests in which the items are segregated.

Simplicity of response. The typical group test requires a very simple response. It may involve underlining a word or making a mark upon a drawing — rarely anything more complicated than this. In some cases, such as the completion tests, it involves writing a word in a blank. In the search for tests which should demand only very simple response, psychologists have devised certain forms which meet these requirements. Among these are the *yes* and *no* tests, the multiple-answer tests, completion tests, and the cross-out tests. In general, these tests have proved very useful; but whether all of them are successful in requiring thoughtful consideration on the part of the person being examined is perhaps a question. The question is most pressing in reference to the *yes* and *no*, or right and wrong tests. It is commonly recognized that these tests tend to encourage guessing, since it is possible to obtain a correct

score in half of the cases by guessing. In fact, the answers to these tests are commonly subjected to a correction by deducting from the number of correct answers a number equal to the number of wrong answers. Doubt has been cast upon this procedure by recent investigations, and it seems reasonable to say that tests or scales which contain tests of other sorts than the *yes* and *no* type of test are so far to be preferred.

Ease and definiteness of scoring. The publishers of most of the prevailing group tests furnish with them stencils or other means by which scoring can be quickly and easily done. In fact, most of these tests can be scored as well by an accurate clerk as by a psychologist. The estimate of the amount of time required to score the test should be considered an important item, in addition to the price, in determining the cost of a testing program.

A device which makes possible very easy and rapid scoring is the Clapp-Young Self-Marking Device, as used in the Henmon-Nelson Test. In this system, test items are printed on the outside of two sheets, the edges of which are glued together to form a sealed folder. Inside the folder are printed squares corresponding to the correct answer spaces of the test items. The inside surfaces of the sheets are provided with carbon ink so arranged that the marks made by the testee are reproduced on the reverse side of the sheet inside the folder. Thus, the number of correct or incorrect responses can quickly be found by unsealing the folder and counting the carbon copy crosses which fall inside or outside the correct answer square.

A machine has now been perfected and put on the market by the International Business Machines Corporation which scores electrically tests of the multiple-choice type. To be scored in this machine, tests are printed with spaces for the answers properly spaced and arranged. The test is merely

inserted in the machine and the score read off on a dial. In the future, tests will doubtless be more and more adapted to machine scoring.

Norms. Among the uses which may be made of the scores in a test is the comparison of the scores of an individual or of a group with a standard or a norm which has been established by giving the test to a large number of persons. The use of such norms often raises rather difficult questions of interpretation. For example, if a school is in a poor district, the scores of a majority of the children will probably be below the norm. The question is, How shall this deficiency be interpreted? Furthermore, if the group is in general below or above the norm, and if we represent the individual's score in terms of his relation to the norm, the distribution of the scores will be lopsided. For purposes of internal administration, therefore, it is probably more useful to compare the scores of individuals with the averages or medians of their own group rather than with an outside norm. However, since norms are useful in some cases, the validity of the norm is a question to be taken into account.

The validity of the norm, in general, depends upon the number of cases and upon their selection. The larger the number of cases the more stable will be the norm. The selection of cases should be such that the individuals do not belong predominately to one or another class of the population, but represent the different classes in the same proportion as they exist in the population as a whole. This may refer to location, to social level, or to age. In the case of age, it is very difficult to get properly selected cases for the ages in the adolescent period, because we usually test children in school, and those who remain in the school constitute only a part of the respective age groups. Furthermore, those who have dropped out of school are usually not equal in intellectual ability to those who remain. The

norms of a test, therefore, should be judged with reference to the number of cases which are used in deriving each norm, and with reference to the way in which the cases have been selected. In the chapter on technique we shall have to consider the relation between age norms and grade norms and the significance of these two types.

The use of an appropriate relative or brightness score. We have already spoken of the methods of calculating brightness or relative ability by means of the intelligence quotient or the coefficient of intelligence. Other scores have been used in connection with certain of the scales. For example, in his original group test, the Advanced Examination, Otis used a score which he calls the index of brightness, or I.B. He finds this by comparing the individual's score with the norm for his age, and then adding the difference between the score and the norm to one hundred, or subtracting it from one hundred, in case the difference is a plus or a minus difference. This gives a score which is similar in appearance to the I.Q., and has some relationship with it in meaning, but which is not identical with it. It may be seen from a casual inspection that there is a fundamental difference in principle between the assumption underlying the coefficient of intelligence, for example, and the index of brightness. In the case of the coefficient of intelligence, a difference in scores of the same amount would produce a different coefficient for succeeding years, because the norm increases in amount. In the case of the index of brightness, on the other hand, a given deficiency or excess in score would produce the same index of brightness at different ages. An analysis would show also that the index of brightness involves a different assumption in regard to the mental growth and the distribution of scores than is assumed by the intelligence quotient. The question which is necessary to raise here is

whether the particular index which is used in a given test is justified by the nature of the distribution of the scores which are obtained from it. The literature which describes the test should give evidence that the author has considered carefully the psychological and statistical principles which underlie the form of scoring which he recommends.

Directions for tabulating the results of a test. In some cases the manuals which go with tests present careful directions for plotting the distribution and calculating the various individual scores which are derived from it. In some cases the directions for tabulating and for calculating the scores of individuals are presented in graphic form, which makes the calculation very rapid and easy.

External criteria of the value of the test. The criteria which are here referred to are of a statistical sort, and are derived from the application of the test and the tabulation of the results. The first criterion which is commonly applied has to do with the distribution of the scores. If a test has range of difficulty which is suitable to the different degrees of ability within the group which is to be tested, the scores will be distributed in fair conformity with the normal probability curve. That is, the greatest number of scores will be near the average, and the frequency of the scores will decrease at the same rate above and below the average. The suitability of a test for the various ages or grades to which it is to be applied may be examined in part by tabulating the distribution of the scores, and examining them to see whether they approach the normal distribution.

A second criterion concerns the progression of the average or median score for the successive age groups to which it is to be applied. The average should progress uniformly, at least throughout the ages below the adolescent period. In some cases the average has been found to progress at about the same rate up to middle adolescence, but in the majority

of cases the increment at each age in this period is somewhat less than it is in the preceding period. It is a reasonably safe rule to consider a test unsuited for particular ages if the average score for the successive ages advances either much more rapidly or much less rapidly than for the other ages which are tested.

The final statistical criterion, of course, is derived from the correlations of the test. As we have already seen, the test should show a high degree of correlation with itself when it is repeated, and a high degree of correlation with some outside measure which is assumed to measure the trait in question. Most of the composite group tests which are on the market do not differ very greatly from one another in these respects, but it is at least a mark of the care in which the test has been worked out when the author furnishes the measures of reliability and of validity.

THE CHIEF POINT SCALES FOR THE MEASUREMENT OF INTELLIGENCE

Performance tests

Army Performance Scale Examination. Washington: Division of Psychology, Medical Department, U.S.A., 1918.

Arthur, Grace. *Arthur Performance Scale.* Chicago: C. H. Stoelting Co., 1925.

Cornell, Ethel L., and Coxe, Warren W. *A Performance Ability Scale.* Yonkers-on-Hudson, New York: World Book Co., 1934.

Pintner, Rudolf, and Paterson, Donald G. *A Scale of Performance Tests.* New York: D. Appleton & Co., 1917.

Stutsman, Rachel. *Merrill-Palmer Scale of Mental Tests.* Yonkers-on-Hudson, New York: World Book Co., 1931.

Nonlanguage tests above the primary grades

Brown, Andrew W. *Chicago Non-Verbal Examination.* Chicago: Institute for Juvenile Research.

Goodenough, Florence L. *The Measurement of Intelligence by Drawings.* Yonkers-on-Hudson, New York: World Book Co., 1926.

Greene, Edward B. *Michigan Non-Verbal Series.* Ann Arbor: Psychological Laboratory, University of Michigan.

Pintner, Rudolf. *Educational Survey Tests.* Columbus: Ohio State University.

Pintner, Rudolf. *Pintner's Non-Language Mental Tests.* Columbus, Ohio: College Book Store, 1920.

Porteus, S. D. *Porteus Maze Test.* Chicago: C. H. Stoelting Co., 1924.

Infant and pre-school tests

Bayley, Nancy. *The California First-Year Mental Scale.* Berkeley: University of California Press, 1933.

Buehler, C., and Hetzer, H. *Tests of Mental Development of Children from 1–6 Years of Age.* Leipzig: J. A. Barth, 1932.

Fillmore, Eva A. *Iowa Tests for Young Children.* University of Iowa Studies in Child Welfare, Vol. XI, No. 4. Iowa City: University of Iowa, 1936.

Gesell, Arnold. *Gesell's Pre-School Child Development Scale.* New York: Macmillan Co., 1925.

Goodenough, Florence L., Foster, Josephine C., and Van Wagenen, M. J. *Minnesota Preschool Scale.* Minneapolis: Educational Test Bureau, 1932.

Linfert, Harriette-Elise, and Hierholzer, Helen M. *A Scale for Measuring the Mental Development of Infants during the First Year of Life.* Baltimore: Warwick & York, 1928.

Kindergarten and primary tests

Baker, Harry J. *Detroit Advanced First-Grade Intelligence Test.* Yonkers-on-Hudson, New York: World Book Co., 1928.

Baker, Harry J. *Detroit Primary Intelligence Test.* For Grades II, III, and IV. Bloomington, Illinois: Public School Publishing Co., 1924.

Baker, H. J., and Kaufmann, H. J. *Detroit Kindergarten Test.* Yonkers-on-Hudson, New York: World Book Co., 1922.

Cole, L. W., and Vincent, Leona. *Group Intelligence Test for School Entrants.* Emporia, Kansas: Bureau of Educational Measurements and Standards, State Normal School, 1924.

Dearborn, W. F. *Dearborn Group Test of Intelligence.* For Grades I–III. Philadelphia: J. B. Lippincott & Co., 1921.

Engel, Anna M. *Detroit First-Grade Intelligence Test.* Yonkers-on-Hudson, New York: World Book Co., 1921.

Haggerty, M. E. *Intelligence Examination.* Delta 1, Grades 1–3. Yonkers-on-Hudson, New York: World Book Co., 1920.

Hildreth, Gertrude H., and Griffiths, Nellie L. *Metropolitan Readiness Tests.* For Kindergarten and Grade 1. Yonkers-on-Hudson, New York: World Book Co., 1933.

Kingsbury, Forrest A. *Kingsbury Primary Group Intelligence Scale.* Bloomington, Illinois: Public School Publishing Co., 1920.

Kuhlmann, F., and Anderson, Rose G. *Kuhlmann-Anderson Intelligence*

Test. For Grades I (First Semester), I (Second Semester), II, and III. Minneapolis: Educational Test Bureau, 1933.

Lowell, Frances. "Group Intelligence Scale for Primary Grades," *Journal of Applied Psychology*, III (September, 1919), 215–47.

Myers, Caroline E., and Myers, Garry C. *Myers Mental Measure.* Chicago: Newson & Co., 1921.

Myers, Garry C. *A Pantomime Group Intelligence Test.* Chicago: Newson & Co., 1922.

Otis, Arthur S. *Otis Group Intelligence Scale, Primary Examination.* Yonkers-on-Hudson, New York: World Book Co., 1920.

Otis, Arthur S. *Otis Quick-Scoring Mental Ability Test.* Alpha Test. Yonkers-on-Hudson, New York: World Book Co., 1936.

Philadelphia Mental Ability Test. For Grades 1A and 2B. Philadelphia: Division of Educational Research, School District of Philadelphia, 1938.

Pintner, Rudolf, and Cunningham, Bess V. *Pintner-Cunningham Primary Mental Test.* Yonkers-on-Hudson, New York: World Book Co., 1923.

Pressey, S. L., and Pressey, L. W. *Pressey Classification Test.* Primary Test, Grades I and II. Bloomington, Illinois: Public School Publishing Co., 1921.

Yerkes, R. M., Bridges, J. W., and Hardwick, R. S. *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick & York, 1915.

Tests for intermediate and upper grades

Baker, Harry J. *Detroit Alpha Intelligence Test.* For Grades V–IX. Bloomington, Illinois: Public School Publishing Co., 1924.

Dearborn, W. F. *Dearborn Group Test of Intelligence.* For Grades IV–XII. Philadelphia: J. B. Lippincott & Co., 1924.

Haggerty, M. E. *Intelligence Examination.* Delta 2, Grades 3–9. Yonkers-on-Hudson, New York: World Book Co., 1920.

Henmon, V. A. C., and Nelson, M. J. *Henmon-Nelson Test of Mental Ability.* For Grades III–VIII. Boston: Houghton Mifflin Co., 1931.

Kuhlmann, F., and Anderson, Rose G. *Kuhlmann-Anderson Intelligence Test.* For Grades IV, V, VI, VII–VIII. Minneapolis: Educational Test Bureau, 1933.

McCall, William A. *Multi-Mental Scale.* New York: Teachers College, Columbia University, 1925.

Monroe, Walter S., and Buckingham, B. R. *Illinois Examination.* I. For Grades III, IV, and V. Bloomington, Illinois: Public School Publishing Co., 1920.

National Intelligence Tests. Scales A and B. National Research Council. Yonkers-on-Hudson, New York: World Book Co., 1924.

Otis, Arthur S. *Otis Quick-Scoring Mental Ability Test.* Beta Test. For Grades IV–IX. Yonkers-on-Hudson, New York: World Book Co., 1937.

Otis, Arthur S. *Otis Self-Administering Tests of Mental Ability.* Inter-

mediate Examination, Grades 4–9. Yonkers-on-Hudson, New York: World Book Co., 1922.

Philadelphia Mental Ability Test. For Grades 4A, 4B, 6B and 8B. Philadelphia: Division of Educational Research, School District of Philadelphia, 1938.

Pressey, S. L., and Pressey, L. C. *Pressey Intermediate Classification and Verifying Tests.* For Grades III–VI. Bloomington, Illinois: Public School Publishing Co., 1921.

Tests for high school

Baker, Harry J. *Detroit Advanced Intelligence Test.* Bloomington, Illinois: Public School Publishing Co., 1925.

Henmon, V. A. C., and Nelson, M. J. *Henmon-Nelson Test of Mental Ability.* For Grades VII–XII. Boston: Houghton Mifflin Co., 1932.

Kuhlmann, F., and Anderson, Rose G. *Kuhlmann-Anderson Intelligence Test.* For Grades IX–XII. Minneapolis: Educational Test Bureau, 1933.

Miller, W. S. *Miller Mental Ability Test.* Yonkers-on-Hudson, New York: World Book Co., 1921.

Monroe, Walter S., and Buckingham, B. R. *Illinois Examination.* II. For Grades VI, VII, and VIII. Bloomington, Illinois: Public School Publishing Co., 1920.

Otis, Arthur S. *Otis Group Intelligence Scale, Advanced Examination.* Yonkers-on-Hudson, New York: World Book Co., 1919.

Otis, Arthur S. *Otis Self-Administering Tests of Mental Ability.* Higher Examination, For High Schools and Colleges. Yonkers-on-Hudson, New York: World Book Co., 1922.

Philadelphia Mental Ability Test. For Grade 9B — Junior High Schools. Philadelphia: Division of Educational Research, School District of Philadelphia, 1937.

Pressey, S. L., and Pressey, L. C. *Senior Classification and Verifying Tests.* For Grades VII–XII. Bloomington, Illinois: Public School Publishing Co., 1921.

Terman, Lewis M. *Terman Group Test of Mental Ability.* For Grades VII–XII. Yonkers-on-Hudson, New York: World Book Co., 1920.

Tests for college and adult level

The Army Alpha Examination. First Nebraska Revision. Lincoln, Nebraska: University of Nebraska, 1937.

Baker, Harry J. *Detroit Advanced Intelligence Test.* Bloomington, Illinois: Public School Publishing Co., 1924.

Colvin, S. S. *Brown University Psychological Examination.* Philadelphia: J. B. Lippincott Co.

Henmon, V. A. C., and Nelson, M. J. *Henmon-Nelson Test of Mental Ability.* College. Boston: Houghton Mifflin Co., 1931.

Kuhlmann, F., and Anderson, Rose G. *Kuhlmann-Anderson Intelligence Test.* Grade IX–Maturity. Minneapolis: Educational Test Bureau, 1933.

Thurstone, L. L. *Psychological Examination for High School Graduates and College Freshmen.* Washington: American Council on Education.

Tests for handicapped children

Brown, Andrew W. *The IJR Intelligence Test for the Visually Handicapped.* Chicago: Institute for Juvenile Research, 907 South Wolcott Avenue.

Hayes, Samuel P. *Revision of the Stanford-Binet Intelligence Tests for the Blind.* South Hadley, Massachusetts: Samuel P. Hayes, Mount Holyoke College, 1930.

Chapter VII

TESTS FOR THE ANALYSIS OF MENTAL CAPACITY

We have seen how mental tests developed from single tests into age scales. In the beginning, the purpose which was back of the development of the single tests was the measurement of specific mental capacities. The evidence is that the early psychologists did not have in mind, primarily, the measurement of general mental capacity. The development of tests which would measure general capacity, or general intelligence, as we have seen, arose in two ways. On the one hand, the studies of correlation, such as those which were made by Spearman and his successors, brought to light the fact that the mental processes are interrelated, and prompted the search for mental tests which are closely interrelated, and which may be supposed, therefore, to measure some general or central capacity. We shall see later how this study of correlation was largely influential in the development of our present-day point scales. On the other hand, the age scale of Binet emphasized general mental capacity through the measurement of a composite of mental traits. While Spearman used as his criterion of the significance of a test for the measurement of intelligence its correlation with other tests, Binet used the criterion of age progress. He worked on the assumption that if a mental test gives scores which advance rapidly with age, it represents general intelligence. Mental maturity, that is, was identified by Binet with brightness. We see, then, that the correlation movement and the age-scale movement both directed attention toward general intelligence rather than towards particular mental functions.

1. Test groups

We have now to consider another type of study of mental tests which represented to some degree the earlier interest in the measurement of special mental capacities as well as the measurement of general capacity. About the time of the appearance of Binet's final revision, a number of psychologists were bringing together groups of mental tests which were selected for the purpose of gaining an all-round inventory of the individual's mental capacity. The individual tests were chosen, not primarily because they correlated with one another, or because they showed marked progress with age, but because they were thought to measure certain specific mental traits which it was important to measure. In some cases a composite score in all of the tests of the group was found, but in all cases some attention was paid to the scores of the individual tests, as well as to the composite score. We may call these collective tests *test groups*.

The test groups, as we have seen, have something of the characteristics of the tests of specialized ability, and something of the characteristics of the composite scales for the measurement of intelligence. They tend to develop toward one or the other of these two extremes. If they develop toward the greater specialized measurement of individual mental capacities, they become profile scales. A profile scale is one which keeps distinct the measures of individual traits, and at the same time exhibits them in relationship to one another in an organized pattern. These we shall describe in the latter part of the chapter. The composite scale, on the other hand, disregards the scores in the individual tests and uses merely the composite score of the entire group of tests.

The organization of a group of tests which shall analyze mental capacity as a whole into constituent elements presupposes that such an analysis is possible. It presupposes

the existence of clearly distinguishable and measurable capacities. The efforts to classify mental capacities and to find tests to measure them have proved peculiarly baffling because the investigator has the double problem of finding whether a particular hypothetical capacity is really a distinct capacity, and at the same time of finding a means of measuring it. As a result, our facts and theories in reference to the testing of special capacities are in a confused state. For example, there is vigorous debate as to whether intelligence is a single capacity, along with other specialized capacities, or whether intelligence itself is made up of elements. (See chapter on "The Nature of Ability.") This being the case, we shall survey the efforts which have been made to analyze mental capacity by means of tests, keeping in mind the variation in the presuppositions regarding the lines which such an analysis should follow.

2. The Healy-Fernald test group

We may first describe in some detail a representative example of a test group, and then mention the other chief groups more briefly. The group which is to be mentioned in detail is that which was devised by Healy and Fernald.[1] This test group, like the Binet scale, arose from a practical need. The principal author, Dr. Healy, was the psychologist of the Psychopathic Institute, which was unofficially connected with the Juvenile Court of Cook County, Illinois. His duty, in this position, was to examine the children who were brought to the court, and to endeavor to discover the cause or causes of their delinquency and the mode of treatment which would be most likely to remedy it. It was necessary, in order to make a thorough diagnosis, to conduct

[1] William Healy and Grace Maxwell Fernald, *Tests for Practical Mental Classification.* Psychol. Monog., Vol. XIII, No. 2. Lancaster, Pa.: Review Pub. Co., 1911.

a mental examination, and it was for this purpose that the series of mental tests was collected and organized.

At the time the Healy-Fernald tests were being developed, the 1908 Binet scale was in use. The authors were not completely satisfied with it, however, on two grounds. In the first place, they wished to determine not only the child's general intellectual level, but also the capacities in which he was strong or weak. In the second place, they wished to discover, if possible, what prominent traits might be utilized in devising curative treatment. The reaction of the child to the various tests, therefore, was kept distinct, and his general mental level was determined, not by calculation of a composite score, but by a general summary of his response to the various single tests.

Another characteristic of this group of tests is that the procedure of giving, and particularly of scoring the tests, was not worked out in as great detail as is common in our present-day scales, or as was the case with the Stanford Revision of the Binet scale. These tests, in this respect, resemble the earlier Binet scale of 1905. The authors did not emphasize the objective score which the child made so much as his general behavior and the way he went about the tasks which were set him. In this respect the tests are like those which have been used for many years by the psychiatrist. The Stanford Revision, as we have seen, presents on the contrary a very elaborate and very detailed set of directions for the giving and scoring of each test. The effort is to make the administration of the scale as objective as possible, and to rely principally upon the score which issues from it. Our present scales have gone still further in the direction of making the presentation and the scoring of the test objective or independent of the judgment of the examiner.

We may illustrate the characteristics of the Healy-Fernald group by a list of the individual tests.

LIST OF THE HEALY-FERNALD TESTS

1. *Picture form board.* This consists of a simple picture pasted on thin board, having certain parts cut out by a scroll saw. These parts are to be fitted in the proper places by the children.
2. *Picture puzzle.* Similar to the picture form board, but with a larger number of pieces cut out.
3. *Construction puzzle A.* This puzzle consisted of a frame and five rectangular pieces. These rectangular pieces can be fitted together into the frame.

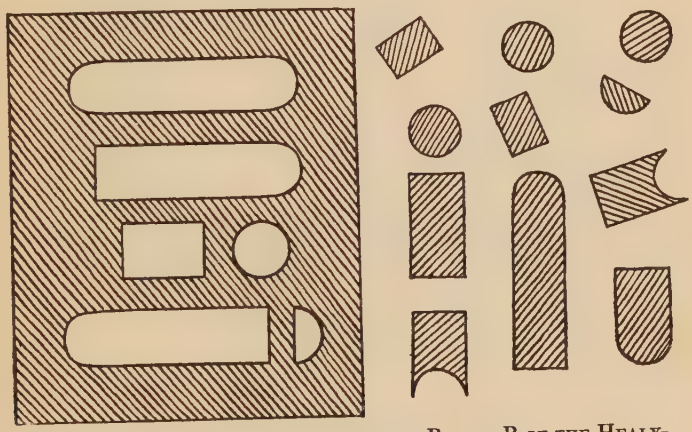


FIG. 7. ILLUSTRATION OF CONSTRUCTION PUZZLE B OF THE HEALY-FERNALD SERIES

(Reproduced by permission of C. H. Stoelting Co.)

4. *Construction puzzle B.* This is similar to puzzle A, but has six spaces to be filled instead of one, and the pieces are of various shapes. (See Fig. 7.)
5. *Puzzle box.* (Fig. 8, p. 174.) This is a box about eight inches square with a glass top. The top is hinged and fastened with a bolt-hook on the front of the box in plain view of the child. This bolt-hook is kept in place by a string which passes to the inside of the box and is hooked over a post. This again is kept tight by another string, which is fastened in another place, and so on. There are five or six steps altogether and the child

may discover how to open the box by tracing the fastenings step by step, and then beginning at the end of the series.

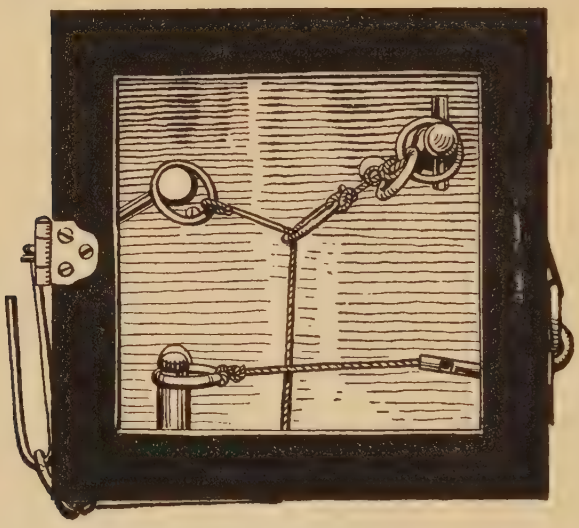


FIG. 8. ILLUSTRATION OF THE PUZZLE BOX OF THE HEALY-FERNALD SERIES
(Reproduced by permission of C. H. Stoelting Co.)

6. "*Aussage*" *or testimony test.* This test consists in showing the child the picture of a butcher shop and making record of the number of things which he is able to report upon after the picture has been removed. The child's suggestibility is also examined by asking him questions about things which were not in the picture and seeing whether he yields to the suggestion.
7. *Drawing.* The child is shown the two Binet figures, each for five seconds. The child is asked to draw the figures from memory.
8. *A simple learning test, by the substitution method.* Nine simple figures are to be substituted for the nine digits. At the top of the sheet is a key containing the simple figures and the digits, each figure to correspond to one digit. Below are series of

rows of the figures to which the child must attach the appropriate digits.

9. *Cross-line test A.* This consists of a simple figure of two cross-lines, in which are inserted the first four numbers. The child is to learn to use the adjacent part of a figure to represent each of the digits.
10. *Cross-line B.* The test is similar to cross-line A, except there are two lines in each direction at right angles to each other, making nine spaces, and there are nine digits instead of four.
11. *Code-test.* This is a complication of the two cross-line tests in which two of the more complicated cross-lines and two of the simpler cross-lines are used together to represent all of the letters of the alphabet. The child learns the code by learning which figures represent the various letters, and then writes a message in the code from memory.
12. *Visual-verbal memory.* Tested by requiring the child to reproduce what he can of a passage which he reads.
13. *Auditory-verbal memory.* A similar test in which the child reproduces what is spoken to him.
14. *Instruction box.* Box in which a small door is fastened by a mechanism which is concealed inside of the box and which can be opened by moving in particular ways certain levers projecting out of the box. The examiner gives the child instructions all at once, and notes the faithfulness and accuracy with which he follows them.
15. *Opposites test.* A series of words which have easy opposites are given to the child and he is required to give the opposite words.
16. *Motor coördination tests.* The child places a dot quickly in as many half-inch squares as he can in thirty seconds.
17. *Handwriting.* The child writes a simple passage in order that the quality of his writing and the mode of his coördination may be observed.
18. *Arithmetic.* The child is given a few simple arithmetic problems.
19. *Reading.* The child is given some simple reading passages.
20. *Checkers.* The child plays a game of checkers with the examiner in order that he may observe how foresighted the child is in his reaction. The test can only be given to children who know the game.

21. *Reaction to moral questions.* The test gives rather incidental information concerning the child's moral attitude.
22. *Information test.* Such questions as, "Who is the President?" "What does Fourth of July celebrate?" "What is the largest city in America?" etc., are given to measure the child's general store of information.
23. There was later added a pictorial completion test which is more elaborate than the picture puzzles in the first two tests.

It will readily be seen that a group of tests of this sort is different in character and in purpose from such a scale as the Stanford Revision. Its aim is not to establish a definite, quantitative measure of the child's intellectual capacity. It is rather to make a more qualitative analysis of his intellectual capacity and of the mode of his reaction to his environment. This analysis is to be made, not simply for the purpose of measuring his various intellectual traits, but also in order that incidental information may be gained concerning weaknesses which may have made it easy to fall into delinquency, and strong traits which may be used in promoting his recovery. The tests are not elaborately standardized, and the scores are not to be used in any precise fashion. For example, there are no age norms presented.

Age norms on these tests were worked out by a later associate of Healy's, Clara Schmitt.[1] The tests were given to children of different ages and the scores made by them were recorded. The scale is not well adapted to a rigid standardization of this sort, however, and it has never been used in the same fashion as our age scales or point scales.

An attempt to work out a definite method of using various tests of this series, as means of diagnosing the special abilities or disabilities which underlie success or failure in the various

[1] Clara Schmitt, *Standardization of Tests for Defective Children.* Psychol. Monog., Vol. XIX, No. 3. Princeton, N.J.: Psychological Review Co., 1915.

school subjects, has been made by Bronner.[1] The attempt, however, is not convincing because it is based on the description of only a few cases. It requires the survey of a large group of cases to demonstrate that defective mastery of reading or arithmetic, for example, is due to deficiency in the particular ability measured by a mental test. It would be necessary to show that a low score in the mental test was uniformly accompanied by deficiency in the subject, and that a high score in the test was uniformly accompanied by success in the subject. This has never been shown, so far as the writer is aware, for any specialized test or any school subject.

There has been some dispute concerning the relative value of the tests which are standardized by elaborate statistical methods, and the qualitative tests such as these of the Healy-Fernald series or the tests which the psychiatrist uses. J. V. Haberman,[2] for example, criticizes sharply the application of rigid statistical methods to the development of mental tests. Haberman insists that standardization is useless in the case of the untrained examiner, and not necessary in the case of the trained examiner. The problem is not to be solved in these terms. Even the trained examiner needs a rigidly standardized test for purposes of making an exact quantitative measure of the child's capacity. This type of measure can also be secured with a fair degree of accuracy, even by the untrained examiner. For the qualitative analysis of the nature of those defects, or of the psycho-physiological conditions for which standardized tests have not yet been devised, the trained examiner is necessary. Few thoroughly standardized tests giving

[1] Augusta F. Bronner, *The Psychology of Special Abilities and Disabilities.* Boston: Little, Brown & Co., 1926.

[2] J. V. Haberman, "The Intelligence Examination and Evaluation," *Psychological Review,* XXIII (1916), 352–79, 383–500.

objective quantitative measures of special intellectual capacities have yet been worked out. Several tests of these hitherto unmeasured traits have recently been devised, however, and no one can say what the limit of their development will be.

3. Other test groups

One of the early test groups, and one which was applied extensively to children prior to our present-day highly standardized tests, is a group which was brought together by Pyle.[1] A distinctive characteristic of Pyle's group of tests is that they were so devised that they could be given to whole classes of children at a time. The aim of Pyle's test was to establish norms for children at different ages, and to measure mental and physical growth by giving the test successively to the same children, or to children of different ages. The test measured, for the most part, rather simple processes of memory and association. One of them, for example, was a simple substitution test similar to that used by Healy and Fernald; another was a test of memory span; a third, of word building; and a fourth of opposites. Pyle gives extensive tables of norms on these tests, but suggests no method by which the scores on the individual tests may be interpreted. No very practical use, therefore, has been made of his group, except that which was made by Pintner.

Pintner adapted a number of Pyle's tests, and added a completion test and an arithmetic test, so as to constitute a group of tests which could be given either individually or together.[2] The scores in the various tests were made comparable by expressing them in terms of percentile scores.

[1] William Henry Pyle, *The Examination of School Children.* New York: Macmillan Co., 1917.

[2] R. Pintner, *The Mental Survey.* New York: D. Appleton & Co., 1918.

The scores thus expressed were then combined, and the combined score expressed in percentiles. This procedure made of the tests a composite scale.

Another group of tests, which is not primarily a composite scale, but which can be combined by the percentile method, is the one devised by Woolley for use in the study of the mentality of working children.[1] These tests were to be given to children who left school to go to work, and also to children who remained in school. The purpose was to study the mental characteristics of the working children, and the effect of their occupation upon their capacity. For this purpose a large variety of mental measurements were made, the aim being to secure, not primarily a composite measure, but a qualitative analysis of abilities. The tests, therefore, included certain psycho-physical tests, as tests of strength of movement, of physical capacity, of rapidity of reaction; and tests of the various mental capacities, such as association, problem solving, memory, imagination, and so on. The tests were scored on a percentile scale.

Another group of tests, which was also devised for a special purpose, was that used at Ellis Island by Knox.[2] This group of tests, because it was to be used with immigrants, was so designed as to avoid the use of language, either in presenting the tests or in responding to them. The group is composed of what are commonly called performance tests. The form board is one illustration of such tests. This consists of a board with openings of various shapes cut out of it, and blocks which must be fitted into the openings. A number of types of form boards were used. Another well-known test of this group is the so-called Knox Cube Test.

[1] Helen Thompson Woolley and Charlotte Rust Fischer, *Mental and Physical Measurements of Working Children.* Psychological Monographs, Vol. XVIII, No. 1. Princeton, N.J.: Psychological Review Co., 1914.

[2] H. A. Knox, "A Scale, Based on the Work at Ellis Island, for Estimating Mental Defects," *Jour. American Medical Assoc.*, XLII (1914), 741–47.

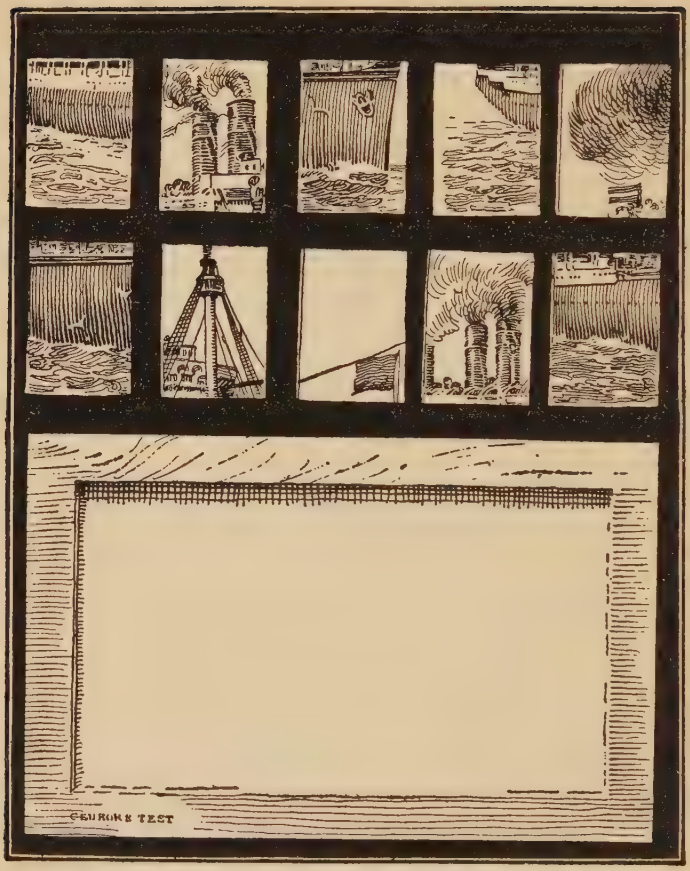


FIG. 9. ILLUSTRATION OF THE SHIP TEST, ORIGINATED BY KNOX AND USED BY PINTNER AND PATERSON, AND LATER BY THE ARMY

(Reproduced by permission of the author and of the publishers, D. Appleton and Company, New York.)

The examiner places four cubes before the subject. He takes a fifth with which he taps the four cubes in a certain order. The examinee must then take the fifth cube and tap the four in the same order. A few of the tests of this group

were borrowed, but many of them were originated by Knox. They have been widely used.

Similar in some respects to the Knox group of tests are the Pintner-Paterson Performance Tests.[1] These tests have been grouped together and called by the authors a scale. There are, in fact, four scales described in the book, but these are supplementary to a detailed and individual description of the particular tests. The tests are independent of language, both in their presentation and in the performance by the children. They are largely of the form-board type. Some of them have been borrowed from Knox, one was a standardization of the Healy Picture-Completion Test, and others have been taken from other sources. There are fifteen in all. An example of this series is the Ship Test, shown in Fig. 9. This test was later used in the Army Performance Scale Examination. Each test was given by the authors to a large number of children, and the results were tabulated in the form of age norms, or standards. The tests are individual rather than group tests — that is, they are given to children singly rather than in classes. The series is useful as an individual test to be given to deaf children, to children who do not understand English, or to those who may have more ability to deal with things than with words.

Groups or collections of tests are still used in clinical examination. The examiner selects such tests as he thinks may reveal significant variations in the abilities of the person he is examining and interprets the responses on them by the liberal use of his own judgment. Tests are often used in this way in case studies in which the aim is to secure a comprehensive picture of the relationship of the various factors or elements in the individual and his environment, for the pur-

[1] Rudolf Pintner and Donald G. Paterson, *A Scale of Performance Tests.* New York: D. Appleton & Co., 1917.

pose of explaining peculiarities in his behavior or emotional life. Collections of tests and manuals for their administration have been provided for the use of the clinical examiner. Examples of such collections are the books by Bronner, Healy, Lowe, and Schimberg,[1] and by Garrett and Schneck.[2]

4. Aptitude tests

The tests which compose the groups mentioned in the preceding sections were selected without any clear conception of the abilities to be measured by the single tests, nor by any objective method of identifying the abilities or validating the tests so selected. The method of factor analysis has not until recently been applied to the selection of tests, and is still not widely applied. Another method has been used for the selection of tests of particular abilities, however, which is more objective than that usually employed, though it is not so analytical as factor analysis. This method is used in making aptitude tests.

The nature of aptitude tests is determined by the definition of an aptitude. In terms of its scope or content, an aptitude is the ability, or collection of abilities, required to perform a specified practical activity. This activity may be that of a vocation or some other of a similar nature. Thus, we speak of aptitude for business, for a trade, for selling, or for a profession. Sometimes the term is used to designate the ability required for the mastery of one of the school subjects, as mathematics, languages, science, or for one of the fine arts, such as music or painting.

The essential characteristic of an aptitude is that it exists prior to training or education in the special field of activity

[1] Augusta F. Bronner, William Healy, Gladys M. Lowe, and Myra E. Schimberg, *A Manual of Individual Mental Tests and Testing.* Boston: Little, Brown & Co., 1927.

[2] Henry E. Garrett and Matthew R. Schneck, *Psychological Tests, Methods, and Results.* New York: Harper & Bros., 1933.

to which it applies, and is not dependent on such training. The aptitude may or may not be native or inherited, but at least it is not the product of special training. Hence, it must be measured before the individual has had special training. Its purpose is to enable us to predict how rapidly and well the individual will acquire ability in the field in question when he does receive training in it.

An aptitude may be psychologically simple or psychologically complex. That is, it may consist of one primary or elementary ability, or in a combination of several. The test in itself does not necessarily reveal the composition of the aptitude. Its composition must be learned by the making of a factor analysis. A factor analysis may be used in designing an aptitude test, but this has not often been done. The usual procedure in designing an aptitude test is to try to assemble a number of tests which will have a high correlation with performance in the practical activity. A number of the tests may be selected because it seems probable on inspection that they will correlate with parts of the activity. They may then be tried out singly, after which the most promising may be assembled into a composite scale. The composite is then finally validated by correlating it with total performance in the activity. If the time comes when primary or elementary abilities can be singled out by factor analysis it may be possible to define aptitudes by determining which primary abilities are essential in the performance of the various practical activities. The design of aptitude tests will then follow a more systematic procedure in place of the empirical process which has prevailed up to the present.

5. Mechanical aptitude tests

Several tests have been constructed for the purpose of measuring mechanical aptitude. The first of these was

devised by Stenquist.[1] It has two tests, one of which is an assembly test. This consists of a number of objects of everyday use which are taken apart and presented to the testee with the instruction to put them together. The other is a pencil and paper test, also containing two tests, each employing the matching technique. The test consists of two groups of pictures of objects. One object in each group is to be matched with an object in the other group. Thus, a wrench is to be matched with a spark plug, a handle with a hatchet blade, a bit with a brace, a seat with a tricycle. In addition, Test II contains an exercise to test the knowledge of names of parts of machines and the ability to trace mechanical relations. This test was validated by correlating the scores with attainment in shop courses. It is better adapted for boys than girls. As a pioneer effort it led the way to the development of aptitude tests in the mechanical field, but it yielded comparatively low correlations with the criterion in the Minnesota study of mechanical aptitudes.

The most elaborate investigation of tests of mechanical aptitude was that made by Paterson and Elliott; it led to the development of the Minnesota Mechanical Ability Tests.[2] These were selected from a large number which were originally tried out. The basis of selection, after satisfactory reliability was secured, was correlation with a criterion, which was success in carrying out projects in high-school shop courses. After a preliminary study seven tests were chosen for final validation. Three of these yielded satisfactory correlations with the criterion, as shown in the table on page 185.

[1] J. L. Stenquist, *Measurements of Mechanical Ability.* New York: Teachers College, Columbia University, 1923.

[2] Donald G. Paterson and Richard M. Elliott, *Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press, 1930.

Test	Reliability Coefficient	Validity Coefficient
Minnesota Assembly, Boxes A, B, C	.94	.55
Minnesota Paper Form Board, Series A and B	.90	.52
Minnesota Spatial Relations, Boards A, B, C, and D	.84	.53

The assembly test is similar to that of Stenquist. It consists of a collection of objects in everyday use which are presented to the testee in disassembled form and which he is required to put together. Among the objects are a bulldog paper clip, an electric bell, an old-fashioned lock, a safety razor, and a small monkey wrench. The paper form board is made up of a series of drawings, each containing a complete figure and a group of parts. Lines are to be drawn in the complete figure to show how the parts fit in. See Fig. 10. The spatial relations test has four form boards, each containing a number of openings of various shapes into which are to be fitted blocks of the same shape. In addition to these tests, Hubbard's Interest Analysis Blank was found to have good correlation with the criterion.

A factor analysis showed that mechanical ability does not consist in a single factor but is probably a combination of several. It is distinct both from intelligence and motor agility. No marked sex differences appear except where practice effects are evident. The lack of correlation with environmental conditions leads the authors to conclude that mechanical ability is largely innate.

The McQuarrie Test for Mechanical Ability includes tests of manual dexterity, maze running, and spatial perception and imagination. It is, according to the findings of Paterson and Elliott, a composite measure of motor ability and mechanical ability.

The Baker-Crockett Detroit Mechanical Aptitudes Ex-

FIG. 10. MINNESOTA PAPER FORM BOARD TEST, SERIES A

(From Paterson and Elliott, *Minnesota Mechanical Ability Tests.* Reproduced by permission of the publisher, The University of Minnesota Press, Minneapolis, Minnesota.)

amination consists of two scales, one for each sex, with norms for grades six to twelve inclusive. Each scale consists of eight tests: 1, knowledge of names of tools; 2, skill of movement as measured by tracing; 3, perception of size; 4, knowledge of the meaning of terms and names of objects; 5, a pencil and paper sorting test; 6, a paper form board test; 7, a test of knowledge of the function of parts of machines; and 8, a test of the ability to trace mechanical relations as represented in drawings. The reliability of the examination, based on 193 pupils, is reported by the authors to be .76. The correlation with ratings by teachers was .64.

It is evident that the conception of mechanical aptitude varies somewhat among the different authors of tests. According to the authors of the Minnesota Mechanical Aptitude Tests it is the ability to perform mechanical operations or tasks and is distinct from intelligence, motor skill, or information. Most of the tests devised by other authors include one or more of these other factors. It is in accordance with the principles of test design to make the test as specific a measure as possible and to use a combination of tests to measure a combination of abilities. Hence, tests of the type of the Minnesota test are to be preferred.

6. Tests of musical aptitude

Attempts to measure musical aptitude have yielded two types of tests, performance tests and pencil and paper tests. The earliest performance tests were those of Seashore.[1] Seashore's series consists altogether of thirty-odd single tests. Their nature may be gathered from the mention of six of the more fundamental ones. These are, first, pitch discrimination; second, discrimination between the intensity of tones; third, the recognition of the relation between

[1] C. E. Seashore, *The Psychology of Musical Talent.* Boston: Silver, Burdett & Co., 1919.

time intervals; fourth, the memory of tones; fifth, discrimination between harmonious and inharmonious combinations of tones; sixth, rhythm. Each of these abilities, along with a large number of others, is considered by the author to be essential to musical performance. He has therefore included them in his scale and devised means of testing them. Seashore does not, however, give us correlation data concerning the relationship between capacity in these individual traits and musical performance in general. On the basis of the scores on the individual tests a profile is drawn. Diagnosis of musical ability is based on this profile.

The Seashore tests were originally built on the basis of a psychological analysis of the elements of musical talent rather than on either factor analysis or correlation with criteria. Such calculations of reliability or validity as have been made have occurred after the tests were put on the market.

The Seashore tests are perhaps the most widely known and used aptitude tests. After they had been used for a number of years at the Eastman School of Music, Stanton [1] studied their prognostic value by comparing the progress and graduation of students who were classified by the tests. She found, for example, that 60 per cent of the "safe" students graduated as against 17 per cent of the "discouraged" group. A correlation study by McCarthy [2] is reported in an article which includes a bibliography and summary of previous studies. This study indicates that the tests of pitch and tonal memory are both reliable and agree with ratings of musical ability but that the tests of time, intensity, and consonance are low in reliability or validity, or in both.

[1] Hazel M. Stanton, "Testing the Cumulative Key for Prognosis of Musical Achievement," *Journal of Educational Psychology*, XXV (January, 1934), 45–53.

[2] Dorothea McCarthy, "A Study of the Seashore Measures of Musical Talent," *Journal of Applied Psychology*, XIV (October, 1930), 437–55.

Similar in general character to the Seashore tests are the Kwalwasser-Dykema Music Tests,[1] which are also on phonograph discs and may be given to a group. These tests are designed to measure ten abilities: 1, tonal memory; 2, quality discrimination; 3, intensity discrimination; 4, feeling for tonal movement; 5, time discrimination; 6, rhythm discrimination; 7, pitch discrimination; 8, melodic taste; 9, pitch imagery; and 10, rhythm imagery.

A number of other tests measure musical knowledge. These are not strictly aptitude tests. One, however, measures aptitude in the form of musical memory. The Drake [2] test, furnished in two forms, contains a number of sequences of melodies, in each case an original and another that is the same or that has been changed in key, time, or pitch. The testee is to indicate in each whether the melody has been changed and, if so, how. This test can be taken by one who does not know musical terms and has not had musical training.

7. *Aptitude for art*

Tests for aptitude in art, as in music, have not been worked out with formal statistical analysis of the abilities which are required for artistic performance. They are based on the judgment of the authors. Three tests will be mentioned, one designed to measure a variety of abilities, and two which measure the ability to judge pictures.

The Lewerenz [3] Tests in Fundamental Abilities of Visual Art are composed of nine tests. One is a test of vocabulary,

[1] Jacob Kwalwasser and Peter W. Dykema, *Kwalwasser-Dykema Music Tests.* Boston: C. C. Birchard & Co.

[2] Raleigh M. Drake, *Musical Memory Test.* Bloomington, Illinois: Public School Publishing Co., 1934.

[3] Alfred S. Lewerenz, *Tests in Fundamental Abilities of Visual Art.* Los Angeles: Department of Psychology and Educational Research, Los Angeles City School District, 1927.

which is hardly an aptitude test. One requires a choice of the most pleasing among four figures which differ in proportion. One requires that the testee indicate where shadows should be placed on groups of simple figures. Three deal with perspective, requiring the testee to mark the lines on a group of figures which are in faulty perspective. One requires the recognition and naming of colors. Two demand drawing, one the reproduction of the outline of a vase and the other the production of original drawings to include groups of dots arranged in random fashion. This last test is intended to measure originality.

The two other tests call only for a discrimination in the artistic merit of pairs of pictures. The older of the two is the Meier-Seashore Art Judgment Test.[1] It has 125 pairs of pictures. One of each pair is the reproduction of an original drawing or painting, reproduced in black and white. The other is an altered picture, changed so as to impair the balance, symmetry, harmony and unity, or rhythm. The testee is to select the better of each pair.

The McAdory Art Test [2] is similar, but it contains four illustrations of each subject instead of two. The process of judgment is the same.

8. Clerical aptitude

Some clerical tests are aptitude tests and some are achievement or proficiency tests. Aptitude tests may be taken before the activity has been learned and may be used to predict learning ability. Achievement tests can be given only after the activity has been learned.

[1] Norman C. Meier, *The Meier-Seashore Art Judgment Test.* Iowa City: Bureau of Educational Research and Service, University of Iowa, 1929.

[2] Margaret McAdory Siceloff and Others, *Validity and Standardization of the McAdory Art Test.* New York: Teachers College, Columbia University, 1933.

A number of them include operations that require quickness and accuracy in finding, sorting, classifying, and arranging printed symbols, such as words, letters, or numbers. Arranging in alphabetical order is one of the procedures commonly included in the tests. Sometimes simple arithmetical operations and spelling are also included. These tests are found to be prognostic of ability in filing and in general clerical operations. It is clear that they do not depend on unlearned activities but only on those which may not have been learned in specific preparation for doing clerical work. Examples of such tests are the Thurstone [1] Examination in Clerical Work, the O'Rourke [2] Clerical Aptitude Test, Junior Grade (Clerical Problems), and the Minnesota Vocational Test for Clerical Workers.[3]

A number of tests have been produced to measure ability in stenography, but they are all tests of proficiency rather than of aptitude. The same is true of tests of ability in typewriting, with one exception, that of Brewington.[4] This test is carried out by means of a typewriter but does not require that the testee shall have learned to use the instrument. In it the individual must press a given key when a given number appears in the opening of a screen. As this key is pressed another number appears, and so on, until the entire series is completed. The test measures the quickness and accuracy with which the testee can learn to associate finger movements with symbols, which is what one

[1] L. L. Thurstone, *Thurstone Employment Tests, Examination in Clerical Work.* Yonkers-on-Hudson, New York: World Book Co., 1922.

[2] L. J. O'Rourke, *Office Employment Tests.* Office Management Series, No. 46. New York: Office Management Association, 20 Versey Street, 1930.

[3] Dorothy M. Andrew and Donald G. Paterson, *Measured Characteristics of Clerical Workers.* Employment Stabilization Research Institute, Vol. III, No. 1. Minneapolis: University of Minnesota Press, 1934.

[4] Ann Brewington, "Tests for Typewriting," *American Shorthand Teacher,* IV (September and October, 1923), 1–5, 50.

does in learning to typewrite. It has been found to correlate well with progress in learning.

9. Aptitudes for academic subjects or fields

Popular educational thought commonly defines aptitudes in terms of the conventional subjects of study or the fields of professional preparation. Tests have been devised to estimate how well pupils are fitted by ability to succeed in several of the subjects or fields of study. These may be classed with aptitude tests.

The broadest test of this type is the Stanford Scientific Aptitude Test, devised by D. L. Zyve.[1] It is composed of eleven parts, each of which is designed to measure one of the fundamental abilities necessary to succeed in scientific pursuits and engineering. These are defined as: experimental bent; clarity of definition; suspended judgment; reasoning; recognition of inconsistencies; recognition of fallacies; induction, deduction, and generalization; caution and thoroughness; discrimination of values in selecting and arranging experimental data; accuracy of interpretation; and accuracy of observation. The composite score discriminates between students specializing in science and engineering, and others, but the validity of the parts of the test is not determined. Other tests of engineering and scientific aptitude have been devised by Thurstone, Roback, and Herring.

Aptitude tests have been produced for subjects at the level of the high school and college. As early as 1918 Rogers [2] developed a test for the prediction of ability in mathematics. The composite finally selected contains two sub-tests which depend somewhat on instruction in algebra

[1] D. L. Zyve, *Stanford Scientific Aptitude Test.* Stanford University, Cal.: Stanford University Press, 1929.

[2] Agnes Low Rogers, *Experimental Tests of Mathematical Ability and Their Prognostic Value.* New York: Teachers College, Columbia University, 1918.

and geometry, and four tests of more general character: a number series test; a spatial relations test; an analogy test; and a language completion test. This is, therefore, partly an aptitude test and partly an achievement test. Later tests distinguish more sharply between the aptitude required in learning a given subject and the result of previous learning. One type of the newer tests consists of actual samples of early learning in the subject itself. This is exemplified in the Luria-Orleans Modern Language Prognosis Test.[1] Another type utilizes tasks which are very similar to those required in learning the subject, as in the Iowa Placement Examinations.[2] These are similar to the tests of engineering or scientific aptitude.

In general, it appears that the aptitude tests which have been developed thus far have been designed and standardized in somewhat the same manner as have the intelligence tests. They represent in part the judgment of a psychologist as to what kind of task will serve to measure a given ability, but this judgment is supplemented and checked by correlation with a criterion consisting of performance or progress in learning in the field in which the ability is exercised. The ability is defined in terms of some practical activity, either vocational or educational, and not in terms of a psychologically pure and independent element of ability, or entity. Whether such entities exist is a question. The answer to this question is sought by means of the process of factor analysis, which will be discussed in another place.

[1] Max A. Luria and Jacob S. Orleans, *Luria-Orleans Modern Language Prognosis Test.* Yonkers-on-Hudson, New York: World Book Co., 1930.

[2] George Dinsmore Stoddard, *Iowa Placement Examinations.* University of Iowa Studies in Education, Vol. III, No. 2. Iowa City: University of Iowa, 1925.

10. Tests of special abilities

There is no general agreement as to what special abilities exist, or indeed whether special abilities exist at all in the sense in which the term is ordinarily used. The use of the term is loose and vague. On the one side, it may be taken to include aptitudes. At the other extreme, it may designate the narrow and independent factors which are assumed to underlie particular performances and which are not common to various performances. Such factors can only be revealed, if at all, by statistical analysis. They are not represented in actual performances which can be identified by inspection. Factor analysis may ultimately enable us to break down abilities into stable and independent factors or elements, and may cause us to abandon the present classification into intelligence, aptitudes, and special abilities. Such an outcome, however, is not certain, and in the meantime it is reasonable to use the categories which are dictated by common sense, psychological insight, and the study of correlation between tests and between tests and performance.

We have already seen that the tests of the earlier period aimed to measure particular, narrow mental functions. We have seen, furthermore, that the results of the correlation between these tests render very difficult the interpretation of the mental functions which they measure. It appears to be necessary, therefore, to carry on a thorough study of the various tests in order that we may be able to identify the functions which are measured by them. The intelligence tests have sidestepped this problem, and have simply set as a criterion the correlation with other tests in general, or with general measures of ability. They have not attempted to find specialized measures of ability.

A good deal of experimentation has been done in the development of single tests. This experimentation differs

from the studies of single tests in the early period, particularly in its emphasis upon tests of a more complex nature and in its elaboration of the test materials.

In the field of sensory tests, there has been little recent experimentation. The technique of testing sensory discrimination is probably thoroughly adequate. The work of Pillsbury, Seashore, and Yerkes and Watson, sponsored by the American Psychological Committee of 1906, still represents the most advanced technique in sensory tests. What we now need is further study of the interrelationships between the tests in the same sensory field in order to determine to what extent they represent general discriminative ability.

In the field of motor capacity there has been greater activity. A study of motor abilities made by Perrin [1] indicates the complexity of the problem. One might suppose motor ability to be a fairly homogeneous affair. We commonly speak of persons as having a high degree or a low degree of manual skill, or skill of movement. It is commonly supposed that the ability to master a skilled operation varies among different persons, and is general in its nature. In order to test this assumption, Perrin gave to about fifty persons three complex motor tests and fourteen simple tests. The complex tests were, first, the Bogardus fatigue test, which requires that a person place a block on a rotating platform; second, a card-sorting test, which requires that cards be sorted into compartments or piles according to some mark upon them; and third, a new motor-coördination test, which requires that a person shall trace simultaneously a square with one hand and a triangle with the other. The fourteen simple tests were of the conventional sort.

[1] F. A. C. Perrin, "An Experimental Study of Motor Ability," *Journal of Experimental Psychology*, IV (February, 1921), 24–56.

Contrary to what we might expect, there was found to be very little correlation between these various motor tests. Even the complex tests, on the whole, correlated only slightly with each other or with the simple tests. In the light of these results, Perrin goes on to discuss the question, "What is motor ability?" He inquires what the evidence indicates as to whether there is such a thing as general motor ability; whether it is a complex or simple unit function; whether it is based on a few general modes of reaction; whether it is closely related to intelligence or to temperament. He does not come to a definite conclusion, except that motor ability is not general, nor a complex of simple unit functions or a few modes of reaction.

There is one condition which is necessary for satisfactory data upon problems such as these, and which was not met in Perrin's study. This condition is the measurement of the reliability of individual tests, or the extent to which each one is consistent. The author's experience has been that sometimes tests which one would suppose to be thoroughly consistent turn out not to be so. Experiments which he has made with a tracing test, for example, prove that this test, although carefully carried out and administered, may have very little reliability. When the reliability was first measured, it was found to be less than .30. By altering the conditions in the administration of the test, it was raised to about .60. If, now, Perrin's tests, as is quite possible, had very low reliability, this may have accounted for the low intercorrelation.

If we include, as a phase of motor capacity, sensori-motor reaction, we find that there has been some investigation both in the field of vocational testing and in the general analysis of this sort of capacity for its own sake. The ability to carry on at the same time and to coördinate a series of parallel reactions was measured by a device which

was first designed for use in the aviation corps of the army.[1] The apparatus provided three different sets of signals, any one of which may be set in operation at any time and which require three sets of responses. It demanded a continuous attentiveness to the three sources of stimulation. This device was built, in the first place, to measure the deterioration in ability to respond which results from a gradual diminution in oxygen content of the air. It was not used to compare the abilities of various individuals with one another, although it might be used for this purpose.

A test which measures continuous reaction is the Pursuit Meter.[2] It is a device composed of various electrical instruments which measures the ability of the individual to adjust his movements to a series of constantly changing stimuli. The object to which the subject is to adjust his movement consists of a spot on a dial which moves behind a line. Through a change in the electric current, this spot moves toward one side or the other side of the line. The individual, by adjusting a rheostat, tries to bring the spot back to the line when it moves away from it. The apparatus measures the accumulation of the errors in the individual's attempt to keep the spot on the line.

Somewhat more complex motor tests or batteries of motor tests have also been devised. Some of them are related to factory operations, such as a machine feeding test devised by Viteles and experimented with by Schultz.[3] Others consist of such operations as assembling. Three such tests are

[1] Knight Dunlap, *Report of Air Medical Service*, p. 300. Washington, 1919.

[2] Walter R. Miles, "The Pursuit Meter," *Journal of Experimental Psychology*, IV (April, 1921), 77–105.

[3] Richard S. Schultz, "A Test for Motor Capacity in the Industries and in the School," *Journal of Applied Psychology*, XII (April, 1928), 169–89.

described by Crockett.[1] They consist in assembling nuts and bolts and inserting them in sockets; packing blocks in a box; and assembling blocks on strips of wood. They might be called aptitude tests since they measure abilities defined in terms of practical activities.

Perception and memory have often been considered to be special abilities in the sense in which the term is here used. In the case of perception there is much difference of opinion as to what should be included. For example, cancellation of letters and card sorting are sometimes called tests of perception. They might, however, be called sensori-motor tests. Because of the difficulty of identifying perceptual ability it is wise to defer listing tests of perception until this ability can be objectively determined.

11. Profile tests

At the beginning of this chapter, it was said that profile tests represent the analytical measurement of capacity. The purpose of test groups is also, as we have seen, the separate measurement of the various capacities. In such a group, however, provision is not made for bringing the measures of the various traits into direct and easy comparison with one another. The profile test represents the development of the test group so that a direct comparison may readily be made. The direction of development is opposite to that of the composite scale, in which the scores of all the individual tests are combined into a composite score.

The idea of the profile test is not new, and various attempts have been made to put it into execution. It has been found more difficult in the execution than in the conception, however. Yerkes, Bridges, and Hardwick proposed, in connection with the description of their point scale, a much

[1] Alexander C. Crockett, "A Measure of Manual Ability," *Journal of Applied Psychology*, XIV (October, 1930), 414–25.

more comprehensive scale which would be of the profile type.[1] These authors suggested a scale of four main divisions. They suggested that the individual parts of such a scale should measure, respectively, receptivity, imagination (including memory), affectivity or feeling, and thought. Each one of these large divisions should contain subordinate divisions. Such a scale as this would be broader than a merely intellectual test. It would involve both the invention and the standardization of a large number of tests which we do not at present possess.

In contrast to this comprehensive plan, there are in existence a few profile tests of a very narrow scope. In fact, they are even less comprehensive than a profile test which covers the range of intellectual capacity. The Downey Will-Temperament Test, for example, which will be mentioned in the chapter on "Tests of Personality Traits" is made up of a series of tests, each one of which measures some aspect of overt behavior. The Seashore music test is another profile scale, which is described under *Tests of musical aptitude.* This scale measures the various constituent capacities which are necessary for musical appreciation and performance.

The general method of the profile scale, then, may be applied to the analysis of any group of mental capacities. It has been very seldom attempted in the general field of intellectual capacity. The first elaborate profile test produced was that of Rossolimo.[2] He reports that the idea of an analytical measurement of the constituents of ability came to him first in 1909. His revised method, including his

[1] Robert M. Yerkes, James W. Bridges, and Rose S. Hardwick, *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick & York, Inc., 1915.

[2] G. J. Rossolimo, *Das psychologische Profil.* Halle: Carl Marhold Verlagsbuchhandlung, 1926.

classification of abilities and description of tests, was described in 1926. The outline of abilities is as follows:

A. Mental Tonus

I. Attention
  1. Intensity of concentration
    a. With choice
    b. With distraction
  2. Span
II. Capability of resisting
  1. Automatism
  2. Suggestibility

B. Memory

III. Perceptual ability
  Visual perception
  1. By recognition
  2. By discrimination
  3. By reproduction
IV. Retention of visual perceptions by the method of recognition
  1. Meaningless linear figures
  2. Meaningless colored figures
  3. Pictures
  4. Objects
V. Retention of elements of speech
  1. Auditory perception of syllables
  2. Auditory perception of words
  3. Auditory perception of words in associative relation to specified syllables
  4. Auditory perception of sentences
VI. Retention of numbers
  1. Auditory perception of numbers
  2. Visual perception of groups of different figures
  3. Visual perception of groups of different marks

C. Higher Association Processes

VII. Simple comprehension
  1. Common and nonsense pictures
  2. Comprehension of series of pictures of sensible and nonsense content

VIII. Combining ability
  1. Putting together pictures containing figures without meaning which have been cut up
  2. Making complicated figures out of several simple parts
IX. Cleverness, the ability to solve simple mechanical puzzles
X. Imaginative ability, as the ability to supply missing parts of pictures, sentences, and words
XI. Power of observation by which the hidden content and peculiar character of an object is recognized

Some of the tests by which these abilities are measured are familiar to students of mental tests and some are new. Each test has ten parts, and the score is the number of parts passed. The scores are plotted on a figure made up by ruling off ten squares opposite the name of each ability. By connecting the points representing the scores a profile is constructed.

The Rossolimo test has not been adapted for use in this country and probably will not be. It is of interest principally because it stands for the notion that an analytical rather than a composite measure of ability is necessary. This belief receives some support from some of the studies in factor analysis, particularly those of Kelley, Holzinger, and Thurstone. However, if primary abilities exist and need to be tested, they should be identified and the tests standardized by more objective methods than those employed by Rossolimo.

Two American tests have been produced recently which seek to provide an analytical measure of abilities. The first is the Detroit Tests of Learning Aptitude, by Baker and Leland.[1] These are intended to serve as a diagnosis of "specific mental faculties." The scheme of diagnosis is represented in the following table. It appears that the designation of the abilities supposed to be measured by the

[1] Harry J. Baker and Bernice Leland, *Detroit Tests of Learning Aptitude.* Bloomington, Illinois: Public School Publishing Co., 1935.

THE TESTS AND SPECIFIC MENTAL FACULTIES

Test	Reasoning and Comprehension	Practical Judgment	Verbal Ability	Time and Space Relationships	Number Ability	Auditory Attentive Ability	Visual Attentive Ability	Motor Ability
1. Pictorial Absurdities	x						x	
2. Verbal Absurdities	x		x					
3. Pictorial Opposites							x	
4. Verbal Opposites			x					
5. Motor Speed and Precision		x						x
6. Auditory Attention Span for Unrelated Words						x		
7. Oral Commissions		x			x	x		x
8. Social Adjustment A	x							
9. Visual Attention Span for Objects							x	
10. Orientation	x	x		x				
11. Free Association			x					
12. Memory for Designs				x			x	x
13. Auditory Attention Span for Related Syllables						x		
14. Number Ability					x			
15. Social Adjustment B	x							
16. Visual Attention Span for Letters							x	
17. Broken Pictures	x			x			x	
18. Oral Directions		x				x	x	x
19. Likenesses and Differences			x					

[1] Harry J. Baker and Bernice Leland, *Detroit Tests of Learning Aptitude, Examiner's Handbook*, page 8. Bloomington, Illinois: Public School Publishing Company. Reprinted by permission.

tests is based on the authors' judgment. They admit that it is "tentative rather than final because knowledge of human behavior is at present both tentative and incomplete." This being the case the question naturally arises whether the attempt to make such a diagnosis is not at present premature.

A somewhat similar plan is represented in the California Test of Mental Maturity.[1] The tests are furnished in four levels, pre-primary, primary, elementary, and advanced. Three tests are designed to measure the simple abilities, visual acuity, auditory acuity, and motor coördination. Twelve other tests are given, classified under memory, spatial relationships, and reasoning. Combinations are made by adding together, respectively, the scores on all the tests, on the language tests, and on the nonlanguage tests. A chart is provided on which the profile of scores may be drawn. The score on each test is translated into mental age by a conversion table, and the profile is expressed in terms of mental age.

The tests for the analysis of mental ability which have been described in this chapter have been designed, for the most part, on the basis of judgment as to how abilities are constituted and how they may be measured, or by correlating test scores with measures of practical activities, the psychological meaning of which is unknown. The attempt to analyze the constitution of ability more precisely is represented in the technique of factor analysis. This technique has also led to the identification of constituents or elements in ability and will doubtless lead ultimately to the adoption of the profile method of representing the relative standing of the individual in the various constituents.

[1] Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs, *California Test of Mental Maturity.* Los Angeles: Southern California School Book Depository, 1937.

Since the matter is approached in a different way than is described in the present chapter its further discussion will be deferred until after the description of factor analysis.

The examples which may have been given may, at least, indicate something of the nature of our problem and of the complications in which it is involved. The investigation of the feasibility of specialized tests is one of the major problems in the future development of this field. It can, probably, best be attacked by an intensive study of certain of the aspects of mental capacity which we are accustomed to regard as fairly distinct and definite, such as motor ability and memory ability. This study may result in the revision of our conception of the classification of mental capacities, or perhaps in the development of an entirely new system of classification. This problem will be discussed further in the chapter which deals with factor analysis.

Chapter VIII

TESTS OF PERSONALITY TRAITS

Intelligence tests have made a large contribution to the analysis of the capacity of pupils to do school work. The scores on intelligence tests have a fair correlation with the achievement of pupils in their courses. The correlation is very far from perfect, however, and there is clear evidence that the comparatively low correlation is due, not simply to errors in the measurement of intelligence and of achievement, but also to the presence of other factors in achievement besides intelligence. An attempt to explain the discrepancy between intelligence and achievement in individual cases frequently brings convincing evidence that this discrepancy is due to some characteristic of the individual's character and temperament rather than to his intellectual capacity. A complete measurement of the factors in school work, or in achievement in general, therefore, must include other traits besides intelligence.

These other traits have been grouped loosely under the general head of *personality*. Personality is not a technical, psychological term, but serves conveniently to cover a number of mental traits which are not intellectual, but which depend, in some measure, upon the individual's native or inherited make-up. As with intelligence, we may attempt to measure these traits without settling the ultimate question of their origin. We need only assume that they are at least partly due to nature as contrasted with nurture, or that they have become so fixed by habit as to be a relatively permanent part of the individual's make-up. Several attempts have been made to classify personality traits, with

diversity in the resultant groupings. Since there is little hope of entire agreement upon any one classification, perhaps the most useful procedure is to follow the one suggested in the tests themselves. It now seems convenient to group the tests under nine heads: will temperament, behavior or conduct, moral judgment or knowledge, social reactions, extroversion-introversion, neurotic tendencies, attitudes and opinions, dominant interests, and composite or miscellaneous. It is impossible even to list the tests that have appeared since the middle twenties, much less to describe each one. The procedure will be to describe one or two tests as examples of each type, and to give a selected list of tests at the end of the chapter.

1. Tests of will temperament

Will temperament designates the characteristics of the individual's overt reactions. Thus, a person may react to the stimuli of his surroundings energetically or weakly. He may, in general, react promptly or slowly. He may be persistent or vacillating. He may proceed cautiously and carefully or recklessly. His ideas may work themselves out into actions easily, or there may seem to be a blocking or obstruction which must be overcome before the action can take place.

In order that we may test such traits as these they must of course exist as general characteristics, and not simply as particular forms of reaction to specific circumstances. A person must be one of quick decision or of slow decision in general, and not simply with a disposition to decide promptly or deliberately in one particular situation. It must be obvious that a person's reactions are affected to some extent by the nature of the circumstances. A person may react very explosively toward another who is weaker than himself, and cautiously toward one whom he fears; but the

assumption is that, underlying these diversities due to the circumstances, there is a general trend which may be discovered by testing all individuals under the uniform conditions which are set up by a standardized test.

The most elaborate test of will temperament is the one devised by June E. Downey.[1] This test was the outgrowth of a prolonged series of investigations of handwriting and of muscle reading. In the study of these forms of behavior Downey was impressed with large variations in the reactions of different individuals, and with the resemblance between their reactions in these specific forms of activity and their general conduct. She found handwriting a very convenient mode of behavior to test, and found that the behavior of individuals when they write, under a variety of conditions, gives a very good indication of their will temperament type. Handwriting, along with two or three additional forms of reaction, then, constitutes the subject-matter of the Downey test.

Downey thinks of the will temperament as based, in the final analysis, chiefly upon two fundamental factors. The first of these is "the amount of nervous energy at the disposal of the individual," and the second is "the tendency of such nervous energy to discharge immediately into the motor areas and innervate the muscles and glands, or, on the contrary, to find a way out by a roundabout path of discharge." [2] The individual's behavior pattern, then, is due fundamentally to the fund of energy which he possesses, and to the openness or the blocking of the paths of discharge. There may exist various combinations of these two conditions.

[1] June E. Downey, *The Will-Temperament and Its Testing.* Yonkers-on-Hudson, New York: World Book Co., 1923. This book contains an account of the theory of will-temperament testing, of the test, and of typical profiles.

[2] *Ibid.*, p. 59.

A more detailed analysis of the will traits brings us to a division into three groups. These are "(1), those of speed and fluidity of reaction; (2), those of forcefulness and decisiveness of action; and (3), those of carefulness and persistence of reaction." [1]

Speed and fluidity of reaction is measured in Downey's test under four heads. We may take these up in turn and describe the tests by which they are measured.

The first characteristic which is measured under this head is *speed of movement.* The test requires the subject to write the words "United States of America" at his ordinary speed. This test assumes that individuals, if left to themselves, adopt a characteristic speed of movement, and that handwriting is a typical activity which fairly represents the individual's general speed of movement. The second part of the assumption is subject to exception in the case of persons of special training or special lack of training, but it is thought to hold for most individuals who have an ordinary education.

The second test is for *freedom from load.* It assumes that some persons habitually work near their maximum level of achievement, and that others are subject to a load or inhibition which keeps their activities at a level below their maximum. The amount of load is measured by comparing the speed of ordinary writing with the speed of maximum writing. The individual is directed to write his name and the words "United States of America" as rapidly as possible. The ratio is then found between the time of the normal writing and the time of the speeded writing. This ratio will, of course, ordinarily be greater than one. A low ratio of 104 to 115 indicates that the normal time is very little greater than the speeded time, and that the individual is characterized by marked freedom from load. A high ratio of from 150 to 220 indicates a great difference between the

[1] *Op. cit.*, p. 62.

two rates of writing, and indicates a great load. Such an individual requires a strong stimulus to induce him to work as rapidly as he is capable of working.

The third measure of fluidity of reaction is a test for *flexibility*. It consists in an attempt to disguise one's writing of the words "United States of America." Some persons are able readily to disguise their handwriting, either through the possession of a dramatic or histrionic type of temperament or through the exercise of ingenuity. The amount of disguise is scored by comparison with a scale of specimens.

The final test for fluidity of reaction measures the *speed of decision*. The individual is presented with a list of twenty-two pairs of opposite traits, and is asked to check the one of each pair which characterizes himself. If he prefers he may grade himself on the two traits instead of merely checking the one which is characteristic. Examples of the pairs are:

careful, careless
cautious, daring
ambitious, unambitious
punctual, tardy

The purpose of the test is not at all to determine whether the person rates himself accurately, but only whether he decides promptly or deliberates long. Great differences between individuals in performing this simple task are found. The interesting feature of the situation is that an individual usually finds good and sufficient reason for reacting as he does, whether rapidly or slowly. The individual does not realize that his mode of reaction is an expression of his individual temperament, and that another person may justify an entirely different type of reaction with as valid reasons as the one he gives himself.

The second group of four traits measures *forcefulness or decisiveness of reaction*. A person who stands high in these traits may be described in a general way as an aggressive

individual. The first test measures what is called *motor impulsion.* This term designates the amount of energy which is behind one's actions. One's movement may be rapid or slow, vigorous or weak, and yet these differences may not represent accurately the amount of force behind the action. This is because the action may be inhibited or blocked in some fashion. The amount of motor impulsion is determined by setting the conditions so that the action will be free from inhibiting factors, or so that the action will be more spontaneous than usual. This is done by making it automatic; that is, by diverting the attention of the individual from what he is doing. The individual is required to write, first, with his eyes closed; second, while counting by threes with his eyes open, and again with his eyes closed; and third, by writing while he is counting the taps of a pencil by twos. If the size of the writing under these conditions is greater than one's ordinary writing, a degree of motor impulsion above the average is indicated. If the writing becomes smaller, a low degree of motor impulsion is indicated. Increase in speed over the normal also indicates high motor impulsion, and decrease in speed a deficiency in motor impulsion.

The second test of this group measures the *reaction to contradiction.* In the early part of the test period the individual is asked to make a purely arbitrary choice between two envelopes. The envelopes are then put aside while the other tests are given. In the present test he is asked to state which envelope he chose. The examiner then contradicts his statement. The mode of his reaction to this contradiction is the basis of the rating. For example, if the subject throws the burden of proof upon the examiner, or suggests that the examiner is in error, or exhibits angry or suspicious behavior, he is scored *ten* in reaction to contradiction. If at the other extreme he makes some such remarks as "You

fooled me that time," or gives in when the examiner says, "Are you sure? I thought it was —" (naming the opposite envelope), or says that the envelope is forgotten, he is given the lowest score, namely *one*.

The third trait of this group is *resistance to opposition*. This is measured by having the subject write blindfolded and then placing an obstacle in front of his pen and noting his reaction. Very strong resistance to opposition, for example, is represented by exerting strong pressure against the obstacle, maintaining the writing at its initial level by a firm, strong stroke, and usually with enlarged characters, the subject requiring no urging. The lowest grade is given to one who shows absolute passivity in spite of urging. A typical remark is, "I can't," or, "How can I when you stop me?" Here, again, the individual is sure to justify his reaction, of whatever type it may be, but the reaction is due, not to his judgment as to what he should do, but to his temperamental characteristic.

The last trait of this group is *finality of judgment*. At the end of the examination the individual is given the pairs of traits which were presented to him at the beginning and asked to make any changes which he wishes in his rating of himself. The degree of finality of judgment is measured by the shortness of time which the individual requires for this rechecking. If he is very well satisfied with his original judgment he takes a short time. If, however, he has a disposition to revise his judgment, he takes longer time.

The last group of four traits represents *carefulness and persistence of reaction*. *Capacity for inhibition*, which may perhaps be regarded as the basis of control, is measured by requiring the individual to slow down his writing. He uses the same phrase as before, and is instructed to write it just as slowly as possible and still keep the pencil moving. He is told, "Some people take thirty minutes to write the phrase.

Do not enlarge your writing." For some persons this is a tremendously irritating task. They fly to pieces and find it apparently impossible to comply with the direction. A score of *five* is given to a person who can devote about two minutes to the task. A score of *ten* is given to one whose time is longer than eight minutes and fifty seconds, and a score of *one* to a person who cannot take more than twenty-six seconds.

The second trait is *interest in detail.* The individual is asked to copy a particular specimen of handwriting, first, as exactly as possible, taking all the time he wishes, and second, without as great emphasis upon exactness and at the individual's own natural speed. The degree of interest in detail is measured first by the accuracy of the imitation, and second by the excess in time taken in careful imitation without special instruction.

The third test in this group measures *coördination of impulses.* The individual is required to write rapidly the phrase "United States of America," on a line about one and one quarter inches long. His score depends upon the degree to which the rapidity of the writing approximates the former speed of writing, and the completeness with which the individual keeps within the line. The successful individuals are the ones who can keep in mind both the requirement of speed and the limitation of extent. The ones who fail neglect either the one or the other of these two requirements.

The last test is called *volitional perseveration.* In the test for flexibility the individual is directed to practice the disguise of his handwriting as long as he wishes on the back of the sheet. He is instructed "Take all the time you wish and do your best." The amount of time taken may vary from twenty-five seconds to fourteen and one half minutes. The time in this exercise is taken as a measure of one's natural persistence.

The scores on the various parts of the will temperament tests are represented in the form of a profile. This profile shows at a glance the traits in which the individual scores high, and those in which he scores low, and enables one to judge of the general character of the individual's will temperament. A specimen profile is shown in Fig. 11. It will be noticed that each of the tests is scored on a scale from zero to ten.

The individual whose profile is before us is apparently low in speed and fluidity of reaction, represented in the first four traits. He is very slow in speed of movement and in speed of decision, is characterized by considerable load, and is not very flexible in reaction. The writer happens to be well acquainted with the individual and can testify that the record of the test in these respects is entirely correct. In two of the second group of traits the individual scores high and in two low. There is a small amount of motor impulsion, and not very vigorous reaction to contradiction. On the other hand, the individual pursues his course of action vigorously when he meets opposition, and holds rather tenaciously to his judgments when he has once made them. He may revise them when contradicted, but is not inclined to question them spontaneously. In the last group also, there are two high and two comparatively low records. Coördination of impulses is the lowest of this group and volitional perseveration is the highest. Interest in detail is also relatively high, and the ability to inhibit reactions somewhat under the average. The high rating on volitional perseveration is certainly characteristic of this individual. This is probably his outstanding trait. Lack of good coördination of impulses is not as evident. The profile, as given by the test, agrees very closely with one which was based upon the estimate of close acquaintances, the correlation being about .65.

Because of the common failure to find much correlation between the scores on the Downey test and the ratings of judges, the prevailing opinion among psychologists is that the test does not measure any real characteristics of the personality.[1] A recent study, however, suggests that this

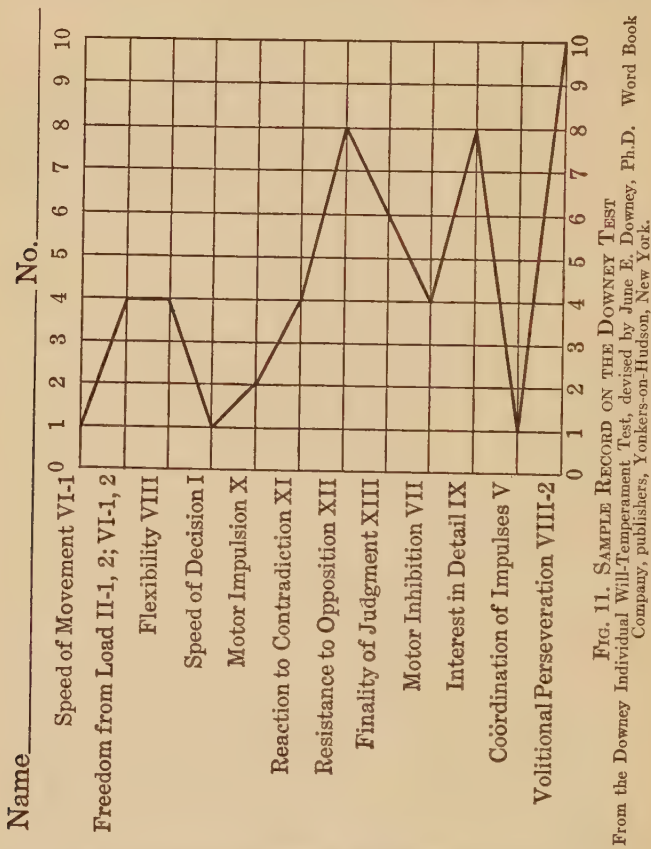


FIG. 11. SAMPLE RECORD ON THE DOWNEY TEST
From the Downey Individual Will-Temperament Test, devised by June E. Downey, Ph.D. Word Book Company, publishers, Yonkers-on-Hudson, New York.

[1] Percival M. Symonds, *Diagnosing Personality and Conduct.* New York: Century Co., 1931.

view may need to be revised. The author, in collaboration with Horatio H. Newman and Karl J. Holzinger,[1] gave the test to nineteen pairs of identical twins who had been separated in infancy. In five of the pairs the profiles of the two individuals were nearly identical, ten are similar in a considerable part, and only four are widely different. One of the most striking cases of close resemblance is illustrated in Fig. 12. That such a resemblance would happen by chance is almost unthinkable. That it could happen in five cases out of nineteen is entirely so. It is fair to conclude then that the Downey test does measure fundamental characteristics, though it may be difficult to identify them.

2. Tests of behavior or conduct

Tests of behavior differ from tests of will temperament in that they measure conduct, that is, behavior which has ethical or moral significance. Another way of describing conduct is to say that it is behavior which is considered by society to have crucial social importance and concerning which standards have been developed which are enforced by various kinds of social pressure, extending from praise and reward to ostracism and punishment. Tests in this field of morals are of two kinds, those which measure conduct directly and those which measure moral or ethical judgment or knowledge. The present section deals with tests of conduct.

The first set of tests which depends on behavior is that devised and used by Voelker. Voelker's tests were designed to measure trustworthiness. They set up situations in

[1] Horatio H. Newman, Frank N. Freeman, and Karl J. Holzinger, *Twins: A Study of Heredity and Environment.* Chicago: University of Chicago Press, 1937.

which the individual has an opportunity to act in an untrustworthy manner. The moral character of the test was disguised and the attempt was made to determine what the reaction of the individual would be when he did not realize that he was under scrutiny. The tests were represented to be mental tests, or tests of mental ability.

Voelker used two series of tests. The first series was

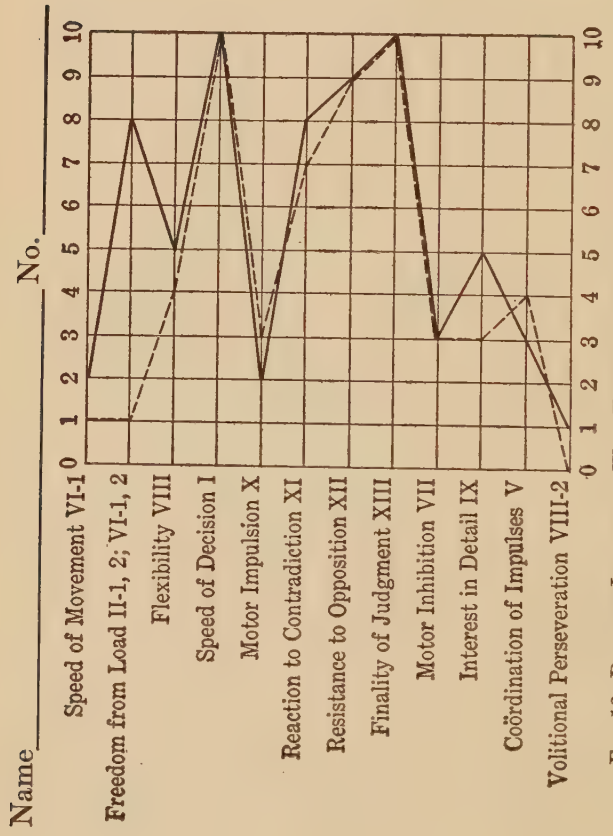

Name
No.
Speed of Movement VI-1
Freedom from Load II-1, 2; VI-1, 2
Flexibility VIII
Speed of Decision I
Motor Impulsion X
Reaction to Contradiction XI
Resistance to Opposition XII
Finality of Judgment XIII
Motor Inhibition VII
Interest in Detail IX
Coördination of Impulses V
Volitional Perseveration VIII-2
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10

FIG. 12. DOWNEY INDIVIDUAL WILL-TEMPERAMENT TEST PROFILE OF A PAIR OF IDENTICAL TWINS WHO WERE SEPARATED IN INFANCY

given to several groups of boys at the beginning of a period of seven weeks, and the second series to the same groups of boys at the end of this period. Two of these groups had received special training and instruction in trustworthiness as members of Boy Scout troops. The second series was aimed to measure the same traits as the first, but were somewhat modified in form. The following is the list of the tests:

Series I.

1. *Overstatement.* Among other questions the boy is asked if he received 95 in his arithmetic.
2. *The M and N suggestibility test.* This test was given as a group test and is adapted from one of the tests of the Downey Individual Will-Temperament Tests. It is aimed to determine whether the boy will stick to his remembrance of a simple fact in the face of a contrary suggestion.
3. *The "Let-me-help-you" test.* This is given as a group test. The child is shown some puzzles and is told to work them without help. An examiner comes in later with some additional puzzles and offers to help the children on the first ones.
4. *The borrowing-errand test.* The child is instructed to borrow something and is told by the leader to have it returned promptly. His score is based on the faithfulness with which he carries out this instruction.
5. *The purchasing-errand test.* The child is given over-change to see whether he will return it or keep it for himself.
6. *The tip test.* The child is given a tip for some trifling favor.
7. *The push-button test.* The boy is told to push an electric button every two minutes by the clock for a certain number of times. The faithfulness with which he does this is kept account of.
8. *The crossing 'a' test.* This is the first part of a group test. The individual is set to crossing out *a*'s in two types of material, one uninteresting and one interesting.
9. *The Pintner profile test.* A group test. The boy is directed to perform this test with his eyes shut. It can only be performed by chance without looking at it. The score is based

upon the proportion of times which the examinee reports that he has done the test correctly.

10. *A group tracing and opposites test.* On the first sheet of a folder, a boy writes the opposites of a list of words. His responses are recorded without his knowledge on wax paper underneath. Later he is put in a situation in which it is possible for him to make correction in his responses and thus unfairly raise his score.

SERIES II.

1. *Overstatement.* Similar to Series I, No. 1, except that different statements are used.
2. *Truthfulness and suggestibility.* Aims to determine whether the individual checks up and contradicts a false statement when he has the facts at his disposal.
3. *Receiving help.* This is a group test in which a series of problems are presented on the first page of a booklet, and answers are given on the back. Some of these answers are wrong so that if the individual copies them, he can be detected.
4. *Reliability.* The boy is directed to deliver a letter and see that it is answered.
5. *Honesty.* A letter is mailed to the boy containing twenty-five cents, which was obviously sent by mistake.
6. *Taking a tip.*
7. *The push-button test.*
8. *The 'a' test.*
9. *The cardboard test.* This is similar to the profile test in its aim. A cardboard containing five circles is presented and the subject is instructed to touch them in turn with his eyes shut. If he reports that he has touched them all he is marked as having failed.
10. *A completion test.* Similar in nature to the tracing and opposites test of Series I.

The evidence which is presented in Voelker's study seems to indicate that the tests have some degree of validity as measures of moral reaction. The groups were judged as to their trustworthiness by teachers and other adults who were acquainted with them. The correlations were found

between the ratings of these judges, and also between the judges' ratings and the test. In some cases the correlations were low, but they were in all cases positive. Correlations between the ratings of the judges and the tests were sufficiently high to indicate a considerable agreement between them.

The most elaborate set of tests of conduct and the most extensive experimentation with them was described in the report of the Character Education Inquiry, carried on by Hartshorne and May, and others.[1] The tests of honesty are the most extensively developed, but some were also devised for measuring service, inhibition, and persistence. Many tests by Maller and others are published by the Association Press, New York.

While not strictly a test, the Haggerty-Olson-Wickman Behavior Rating Schedules should be mentioned here because of their extensive use. As the title indicates, it is not a direct measure of conduct, but a rating scale. It is made up of two main parts, Schedules A and B. Schedule A is a list of fifteen behavior problems, on which are checked those which appear in the child being rated. This then is a check list which is to be marked in terms of frequency of occurrence. Schedule B has four divisions, dealing with intellectual, physical, social, and emotional traits. There are thirty-five items, on each of which the child is to be rated on a graphic rating scale. These may be illustrated by Item 20. The numbers underneath the descriptive phrases indicate the rank order of the average scores of the groups receiving the rating in question. As in the present

[1] Hugh Hartshorne and Mark A. May, *Studies in Deceit.* New York: Macmillan Co., 1928; Hugh Hartshorne, Mark A. May, and Julius B. Maller, *Studies in Service and Self-Control.* New York: Macmillan Co., 1929; Hugh Hartshorne, Mark A. May, and Frank K. Shuttleworth, *Studies in the Organization of Character.* New York: Macmillan Co., 1930.

20. How does he accept authority?

Defiant	Critical of authority	Ordinarily obedient	Respectful, Complies by habit	Entirely resigned, Accepts all authority
(5)	(4)	(3)	(1)	(2)

case, it appears that neither extreme rating is received by the best behaved children. It depends on the actual behavior of the children who are given the various ratings, not on speculative judgment.

3. Tests of moral judgments or knowledge

The significance of tests of moral judgments and knowledge has been questioned on the ground that, as shown both by experience and by statistical inquiry, knowledge and conduct may not agree. That is, the ability of the individual to say what is right or what should be done under given circumstances may not agree with what he will actually do if confronted with these same circumstances. Hartshorne and May [1] give the correlation between knowledge and conduct based on the average of eight tests as follows:

TABLE X. CORRELATIONS BETWEEN MORAL KNOWLEDGE AND CONDUCT FOR INDIVIDUALS AND GROUPS
AVERAGE OF EIGHT TESTS

	Individual	Group
Behavior C (undesirable)	− .25	− .44
Behavior A (undesirable)	− .13	− .15
Behavior H (desirable)	+ .23	+ .53

[1] Hugh Hartshorne and Mark A. May, "Testing the Knowledge of Right and Wrong," *Religious Education*, XXI (February, April, August, October, and December, 1926), 63–76, 239–52, 413–21, 539–54, 621–32; XXII (May, 1927), 523–32.

The correlations with undesirable behavior are negative, as we should expect. The correlations between the knowledge and the behavior of groups are higher than between individuals, which may indicate that group opinion influences group conduct more than individual opinion influences individual conduct. This suggests that the individual's conduct may depart from his knowledge because his conduct is influenced by his feelings, emotions, or self-interest, and also because his expression of opinion may not always represent his real opinion or conviction. We must, therefore, be on our guard against undue reliance on knowledge or expression of opinion as a basis for predicting conduct. However, it is useful to know what a person believes or knows because, if for no other reason, one cannot act on knowledge he does not possess.

One of the earliest of the ethical discrimination tests is also one of the most comprehensive, the Kohs Ethical Discrimination Test. It includes a number of types which have been employed by others. The parts of the test are as follows: 1. Social Relations. The pupil is asked to check the best answer to a number of questions, as for example, "If you have broken something which belongs to someone else you should (a) buy a new one, (b) feel sorry, (c) hide the pieces." 2. Moral Judgment (from Pressey). One is required to indicate which one of each of twenty-five groups of offenses is the worst. 3. Proverbs. Requires the designation of one of three statements of the meaning of proverbs. 4. Definitions of Moral Terms. Definitions of ethical terms in the form of multiple choice. 5. Offense Evaluation. One is to indicate whether the proper thing to do in reference to, say, burglary, is to praise, do nothing, scold, put in jail, put in prison, or kill. 6. Moral Problems (adapted from G. G. Fernald, "Defective-Delinquent Class Differentiating Tests"). Similar in form to Test 1, but designat-

ing reasons why one should not do various things, such as setting fire to a house.

Some of these tests may be criticized because they deal with questions which are theoretical or remote from the experience of the child. Two later tests present situations that are like those the child has to face. One, by Tomlin, is called "The Best Thing To Do." There are forty-five items like the following: "Anne forgot to return Mary's favorite book. It would be a good idea for Mary (a) to tell the other girls that Anne is not honest, (b) not to play with Anne again, (c) to ask Anne to return it, (d) not to lend Anne another book."

More elaborate is the Test of Knowledge of Social Usage by Strang, Brown, and Stratton. It deals with correct usage in reference to 1, table manners; 2, taste in dress and appearance; 3, good manners for guest and host; 4, good form in walking with people; 5, showing respect and consideration for others; 6, good form in talking with people; 7, acting in relation to a group; 8, showing respect for property; and 9, how to act at performances and games.

4. Social reactions

Social reactions are tested by determining how one reacts to other persons. There are a number of types under this heading. Two tests seek to measure social maturity. Furfey's scale deals only with the development of boys. It might be classified as a test of interests, but the types of interests are largely social. It asks the boy to indicate his choice among "Things To Do," "Things To Be When You Grow Up," and "Things To Think About." The Vineland Social Maturity Scale consists of 117 items of behavior arranged in maturity sequence. The person using the scale is to designate whether or not the individual being rated performs the acts described. They range from such acts

as balancing head, standing alone, eating with a spoon (which are hardly social in character) to going to bed unassisted, combing or brushing hair, caring for self at table, making minor purchases, playing difficult games, buying own clothing necessities, and, finally, performing expert or professional work and advancing the general welfare.

A number of tests are for the purpose of measuring particular social reactions. An example of these is the Allport and Allport test of ascendancy and submission. This test is in two forms, one for men and one for women. In each the person is asked to indicate how he would react in each of thirty-three imagined situations. The following example will illustrate: "Some one tries to push ahead of you in line. You have been waiting for some time, and can't wait much longer. Suppose the intruder is the same sex as yourself, do you usually (a) remonstrate with the intruder; (b) 'look daggers' at the intruder or make clearly audible comments to your neighbor; (c) decide not to wait, and go away; (d) do nothing?"

It is assumed in such tests that an individual's reactions in social situations are somewhat generalized, that samples will indicate the nature of one's general reaction, and that one will report with sufficient faithfulness and accuracy what his reaction would be. Other tests of similar nature are listed at the end of the chapter.

Finally, a comprehensive social test is offered by Moss. This test is as much a test of ability as of personality, but it is included here because it belongs as much in one class as the other. It is designed to measure 1, judgment in social situations; 2, memory for names and faces; 3, recognition of mental states from facial expressions; 4, observation of human behavior; 5, social information; and 6, recognition of the mental state of the speaker. Statistical

studies find a close correlation with general intelligence and seem to indicate that the test measures intellectual more than social factors.

5. Tests of extroversion-introversion

The classification of people into extroverts (or extraverts) and introverts was made by the psychiatrist, Jung. An introvert is a person who, in making adjustments, turns his attention inward, tends to withdraw from the external world and to occupy himself with his own thoughts. The extrovert, on the other hand, turns his attention toward the external situation and attacks it vigorously, sometimes with more energy than the situation warrants. Jung himself classified a person after prolonged and repeated interviews. Many psychologists have made selections of symptomatic forms of behavior and put them into the questionnaire or rating form. It may perhaps be questioned whether a satisfactory analysis can be made by the use of these inventories except in the hands of a competent clinician, but in such hands they are useful.

Since the items in the various tests are very similar a few specimens from the Heidbreder scale will suffice. The characteristics of the introvert are described. Those of the extrovert are the opposite. The introvert, for example, limits his acquaintances to a select few, is suspicious of the motives of others, indulges in self-pity when things go wrong, gets rattled easily, day dreams, talks to himself, keeps a diary, is absent-minded, and so on. A few other tests are listed at the end of the chapter.

6. Tests of neurotic tendencies

An inventory to measure neurotic tendencies was first designed by Woodworth during the War. He collected a list of symptoms of neurosis given by psychiatrists and

made them into a questionnaire that could be answered by the person himself or by an acquaintance or observer. Woodworth's Psycho-neurotic Inventory was adapted by Mathews for use with children and called the Personal Data Sheet. A collection of items from various sources was made by Thurstone for his Personality Schedule. A revision of the Woodworth inventory appears in the Woodworth-House Mental Hygiene Inventory.

These illustrations of the kinds of items included in such inventories are taken from the Thurstone Personality Schedule. The underlined answers are the ones which indicate a neurotic disposition. It will be observed that there is a similarity between these answers and those that indicate introversion. That there is a similarity has been confirmed statistically.

yes no ? As a child did you like to play alone?
yes no ? Do you feel that life is a great burden?
yes no ? Do you find it difficult to get rid of a salesman?
yes no ? Do you laugh easily?
yes no ? Do you usually get turned around in new places?
yes no ? Are you afraid of falling when you are on a high place?
yes no ? Are your feelings easily hurt?

Two further tests may be mentioned because they classify the person's responses according to the different kinds of situations he meets. The Adjustment Inventory by Bell deals with adjustment to home, health, social situations, and emotional experiences, respectively. For example, the question, "Did you ever have a strong desire to run away from home?" obviously refers to home adjustment, "Do you take cold rather easily from other people?" to health, "Do you enjoy social gatherings just to be with people?" to social adjustment, etc. Such an inventory is clearly

broader in scope than one dealing only with neurotic tendencies.

The Symonds and Block Student Questionnaire is for young people. The situations are classified under the heads of (1) Curriculum, (2) Social life, (3) School, (4) Teachers, (5) Other pupils, (6) Home and family, and (7) Personal.

7. Attitudes and opinions

The tests of attitudes and opinions follow in the main the plan worked out by Thurstone for developing attitude scales. He assumes that attitudes on controversial questions are the expression of affects or feelings. He assumes further that the feeling concerning such a question will determine one's reaction to the several statements which may be made pro or con with reference to the question and that the statements themselves may be scaled with reference to the extent to which they favor one or the other side of the controversy. The test then consists of a list of scaled items, and the person taking it is asked to indicate any of the statements with which he may agree. His score is the average scale value of these statements. Examples of statements in the Droba Scale of Attitude toward War are: "The benefits of war far outweigh its attendant evils"; "There is no progress without war"; "It is our duty to serve in a defensive war"; and "He who refuses to fight is a true hero."

8. Tests of dominant interests

Attitudes and interests are alike in that they are based on feeling toward an object, idea, or activity. The difference, when a distinction is made, is that attitudes are feeling reactions toward objects and ideas, and interests are feeling reactions toward activities. However, the distinction is not a sharp one.

Tests of interests are usually somewhat covert in their approach. That is, the purpose is, to some extent, disguised. The disguise is accomplished by attempting to discover interest in some broad, general line of activity, such as a vocation, by getting at the interests in particularized activities which are carried on in the pursuit of the vocation. The interest is discovered by having the individual respond to a questionnaire. It is believed that he can respond more accurately and will respond more sincerely about particularized than about more generalized activities.

Tests of interest vary from those which cover the whole range of interests to those which are narrowed down to a particular field. The majority of tests have for their purpose the diagnosis of vocational interests, some covering all kinds of vocations and some only one.

The most general test of interests is Allport and Vernon's Study of Values. This attempts to determine the relative strength of the individual's interests under six general heads. The classification follows that developed by Edward Spranger in his *Types of Men*. It posits six basic interests or motives: the theoretical, economic, aesthetic, social, political, and religious. These are described by Allport and Vernon as follows: "The dominant interest of the theoretical man is the discovery of *truth*. . . . The economic man is characteristically interested in what is *useful*. . . . The aesthetic man sees his highest value in *form* and *harmony*. . . . The highest value for this type (the social) is *love* of people. . . . The political man is interested primarily in *power*. . . . The highest value of the religious man may be called *unity*."

The test consists of two parts. The first part has thirty questions, each calling for a choice between two answers. For example: "At an exposition, do you chiefly like to go to the buildings where you can see (a) automobiles, (b) sci-

entific apparatus or chemical products?" The first answer indicates economic interest with a weight of 2 and the second theoretical interest with a weight of 1. The second part is somewhat similar, each question having four possible answers.

It is not assumed that a person belongs to one type exclusively. Mixed types are recognized. It is found that the profiles of groups of persons in different occupations differ somewhat in the direction in which one would expect.

The more specifically vocational interest tests are represented by the Strong Vocational Interest Blank. This test has eight parts, each of which explores interests in a given area of experience or activity. The fields are, respectively: occupations, amusements, school subjects, activities, peculiarities of people, order of preference of activities, comparison of interest between two items, and rating of present abilities and characteristics. The test was standardized by giving it to people in various occupations, and tabulating their responses on the various parts of the tests. An individual's vocational preference is determined by comparing his responses with those of each occupation in turn. Centers for scoring have been set up by various organizations, among them the Psychological Corporation.

A somewhat similar inventory is that of Stewart and Brainard. The classes of interests are narrower in scope, as illustrated by the titles: physical work, mechanical work, outdoor activity, vocal expression, drawing and art work, leadership, social activity, and so on to the number of twenty. In addition, eight other classes are given of a somewhat different sort. A questionnaire designed especially for high-school students has been prepared by Garretson and Symonds. It deals with interest in occupations, activities, school subjects, job activities, school paper, football team, student activities, and prominent men.

9. Composite and miscellaneous

An early test which has furnished suggestions for a number of later tests is that by Pressey. He calls the test a group scale for investigating the emotions. Pressey's scale makes use pretty largely of experience with abnormal mental attitudes, and emphasizes the pathological emotional conditions. The four tests may be described as follows. The first test aims to discover various special types of unpleasant feeling. In the second test we find an adapted form of the free association test, which again seeks to uncover pathological and criminological attitudes. The third test is an ethical discrimination test, and the fourth test is aimed to discover certain anxiety tendencies. The particular character of these tests may be gathered from more detailed illustrations.

In *Test I* the subject is instructed to cross out every word which is unpleasant. The first two lines are as follows:

1. Disgust, fear, sex, suspicion, aunt
2. Roar, divorce, dislike, sidewalk, wiggle

An analysis of the words indicates that they are so chosen as to arouse different types of fear, or of unpleasant feeling. The types are represented in the first four words of the first list. In each list there is also a neutral word, which is put in as a joker to indicate whether the individual is following the instructions. In each succeeding line the types are represented in the same order, except that they are dropped back one word in the list. In the first list the joker is at the end. In the second list, it is next to the end, and so on. The subject is next instructed to mark the one word in the list which is the *most* unpleasant. The test is scored in terms of the total number of words which are crossed out, which is an indication of one's general emotionality, and the deviation in the words which are marked as being most unpleasant from

those which are most frequently marked by people in general.

Test II. This test consists of twenty-five lines of words such as the following:

1. BLOSSOM, flame, flower, paralyze, red, sew
2. LAMP, poor, headache, match, dog, light

The subject is directed to cross out all of the words in small letters which are connected in his mind with the words in capitals at the beginning of the line. This is a free association test in group form. The aim is to discover pathological trends of association.

Test III gives lists of words representing different types of conduct. This is an adaptation of Fernald's ethical discrimination test. Two of the lists are as follows:

1. Begging, swearing, smoking, flirting, spitting
2. Fear, hate, anger, jealousy, suspicion

Test IV aims to discover anxiety tendencies. As in the first case, the subjects are told to cross out the names of all the things in each list about which they have ever worried. They are also told to draw a circle about the things in each list about which they have worried most. The first two lists are as follows:

1. Injustice, noise, self-consciousness, discouragement, germs
2. Clothes, conscience, heart-failure, poison, sleep

The types of pathological attitudes which are represented in the first list are as follows. Paranoid or suspicion attitude is represented by worry concerning *injustice*; the neurotic attitude by anxiety about *noise*; the self-conscious or shut-in personality by anxiety concerning *self-consciousness*; marking *discouragement* indicates a melancholic or self-accusatory attitude; and marking *germs*, the hypochondriacal attitude. This test, like all the others, is scored in terms of the total

number crossed out, and also in terms of the peculiar choices which are indicated by the words which are in circles.

The scores on all of the four tests are added, and this score is taken to indicate total emotionality. The deviations are then added and the total is taken to express idiosyncrasy in emotion. The author does not, however, emphasize merely these total scores, but emphasizes the desirability of making an analysis of the subject's responses.

Another early test is the Kent-Rosanoff Free Association Test. It consists of one hundred words to each of which the person is asked to respond by giving the first word which comes into his mind. His responses are compared with those commonly given by people in general. Frequent departures from such common responses are regarded as symptomatic of mental aberrations and are followed up in the attempt to discover their significance. A revision of the standardization has been made by Arthur.

A highly individual test, used chiefly in psychiatric examinations, is the Rorschach Ink Blot Test. It is a series of ink blots of fantastic shape, some of them colored. The individual is asked to tell what objects he sees in the blots. His answers are subjected to an elaborate system of interpretation. The test is useful only in the hands of specialists.

A test based on the psychiatric concept of emotional maturity is the Willoughby Emotional Maturity Scale. Emotional maturity may be described as freedom from egoistic or infantile motives or as objectivity of social judgment. There are sixty items, and those are to be checked which characterize the person being rated. He may be rated by somebody else or he may rate himself. His score is the average weight of the items checked. Items which indicate low emotional maturity are: "S. chooses his course of action with reference to his own maximum immediate satisfaction," and "S. characteristically appeals for help in the

solution of his problems." Items which indicate high emotional maturity are: "S. takes a rationalistic view of evil, desires to learn more about the problem and to put into practice what he has learned" and "S. evaluates suggestions without heat, settling the issue upon rational bases, and cannot be persuaded to alter a matured decision except on the basis of new evidence."

Two composite tests will be mentioned. The best known is the Bernreuter Personality Inventory. The method of the test is to score each item according to each of the traits which the test is designed to measure. Each of 125 questions may be answered in terms of "Yes," "No," or "?" Each answer is then assigned a weight according to each trait. In the original test four traits are distinguished, described by the adjectives "neurotic," "self-sufficient," "introvert," and "dominant." In a supplementary standardization by Flanagan two others were added, "self-confidence" and "sociability." To illustrate:

Question	Trait	Weight for each answer		
		Yes	No	?
Do you daydream frequently?	Neurotic	5	− 4	− 2
	Self-sufficient	1	− 1	− 2
	Introvert	3	− 4	0
	Dominant	− 1	1	2
	Self-confidence	3	− 5	0
	Sociability	2	− 3	5

High scores indicate the possession of the trait in high degree except in the case of self-confidence and sociability in which the scores are reversed. The weights for "neurotic" and "introvert" are similar and one may be used to represent both.

The other composite test is the Humm-Wadsworth Temperament Scale. This scale consists of 318 questions to be

answered "Yes" or "No." The answers are classified according to Rosanoff's theory of personality so as to represent seven groups of traits: the normal, hysteroid or antisocial, the cycloid (manic phase), cycloid (depressed phase), schizoid (autistic phase or dementia praecox), schizoid (paranoid phase), and epileptoid. The test is obviously applicable especially to the discovery of individuals who are pathological or verging on the pathological.

10. General comments

The same question may be raised concerning tests of personality as of tests of ability: What are we measuring? Is it sufficient to accept the apparent or common-sense meaning of the tests or must we probe more deeply to get at underlying factors? The method by which basic or underlying factors in ability are sought is factor analysis. The same method has been applied in the field of personality, but the results are not yet conclusive. Whether personality tests will ultimately be designed to measure the factors revealed by factor analysis remains to be seen.

Tests of personality have not yet been found as serviceable for routine use by the teacher or school administrator as have tests of ability. Their meaning in terms of everyday behavior is not so clear. Many of them have turned out to be fairly reliable when given under favorable conditions, but their validity has to be largely taken on faith, and their significance judged on the basis of clinical experience. The labor spent in devising tests has not been matched by equal labor spent in discovering their meaning in the classroom, the playground, or on the job. They yield data on behavior which ought to be of use in teaching and guidance. Their full use will depend on more extensive study of the practical application of the tests.

SELECTED BIBLIOGRAPHY OF TESTS OF PERSONALITY

Will temperament

Downey, June E. *Downey Individual Will-Temperament Test.* Yonkers-on-Hudson, New York: World Book Co., 1921.

Downey, June E. *Will-Temperament, Group Test.* Yonkers-on-Hudson, New York: World Book Co., 1920.

Behavior or conduct

Baker, Harry J. *Telling What I Do.* Bloomington, Illinois: Public School Publishing Co., 1930.

Haggerty, M. E., Olson, W. C., and Wickman, E. K. *Haggerty-Olson-Wickman Behavior Rating Schedules.* Yonkers-on-Hudson, New York: World Book Co., 1930.

Maller, J. B. *The Self-Marking Test.* New York: Teachers College, Columbia University, 1930.

Voelker, Paul Frederick. *The Function of Ideals and Attitudes in Social Education.* Teachers College Contributions to Education, No. 112. New York: Teachers College, Columbia University, 1921.

Moral judgments or ethical discrimination or knowledge of social usage

Kohs, S. C. *Ethical Discrimination Test.* Chicago: C. H. Stoelting Co., 1922.

Lincoln, Edward A., and Shields, Fred J. "An Age Scale for the Measurement of Moral Judgment," *Journal of Educational Research,* XXIII (March, 1931), 193–97.

Strang, Ruth, Brown, Marion A., and Stratton, Dorothy C. *Test of Knowledge of Social Usage.* New York: Teachers College, Columbia University, 1933.

Tomlin, Frank E. *The Best Thing To Do.* Stanford University, California: Stanford University Press, 1931.

Social reactions

Allport, Gordon W., and Allport, Floyd H. *A–S Reaction Study.* Boston: Houghton Mifflin Co., 1928.

Barry, Herbert Jr. "A Test of Negativism and Compliance," *Journal of Abnormal and Social Psychology,* XXV (January–March, 1931), 373–81.

Bogardus, E. S. "Social Distance and Its Origins," *Journal of Applied Sociology,* IX (1925), 216–26.

Furfey, Paul Hanly. "A Revised Scale for Measuring Developmental Age in Boys," *Child Development,* II (June, 1931), 102–14.

Marston, Leslie R. *The Emotions of Young Children.* University of Iowa Studies in Child Welfare, Vol. III, No. 3. Iowa City, Iowa: University of Iowa, 1925.

Moss, F. A., Hunt, T., Omwake, K. T., and Ronning, M. M. *Social Intelligence Test.* Washington: Center for Psychological Service, 1927.

Extroversion-introversion

Conklin, Edmund S. "The Determination of Normal Extrovert-introvert Interest Differences," *Pedagogical Seminary*, XXXIV (March, 1927), 28–37.

Heidbreder, Edna. *Minnesota Personal Traits Rating Scales: Introversion-Extroversion and Inferiority Attitudes.* Chicago: C. H. Stoelting Co.

Laird, D. A. *Colgate Personal Inventory Rating Scales.* Hamilton, New York: Hamilton Republican, 1925.

Marston, Leslie R. *Personality Rating Scale.* Iowa City, Iowa: Iowa Child Welfare Research Station, State University of Iowa, 1925.

Neymann, C. A., and Kohlstedt, K. D. *Neymann and Kohlstedt Introversion-Extroversion Test.* Chicago: C. H. Stoelting Co., 1929.

Neurotic tendencies

Baker, Harry J. *Telling What I Do.* Bloomington, Illinois: Public School Publishing Co., 1930.

Bell, Hugh M. *The Adjustment Inventory.* Stanford University, California: Stanford University Press, 1934.

Symonds, Percival M., and Block, Virginia Lee. *Student Questionnaire.* New York: Teachers College, Columbia University, 1932.

Thurstone, L. L., and Thurstone, Thelma Gwinn. *Personality Schedule.* Chicago: University of Chicago Press, 1929.

Woodworth, R. S., and House, S. D. *Woodworth-House Mental Hygiene Inventory.* Chicago: C. H. Stoelting Co., 1928.

Woodworth, R. S., and Mathews, Ellen. *Personal Data Sheet.* Chicago: C. H. Stoelting Co., 1924.

Attitudes and opinions

Case, A. T. *A Test of Liberal Thought.* New York: Teachers College, Columbia University, 1928.

Thurstone, L. L. (Editor). *The Measurement of Social Attitudes.* Chicago: University of Chicago Press, 1930, 1931.

Watson, Goodwin B. *The Measurement of Fair-Mindedness.* New York: Teachers College, Columbia University, 1925.

Dominant interests

Allport, Gordon W., and Vernon, Philip E. *A Study of Values.* Boston: Houghton Mifflin Co., 1931.

Cowdery, K. M. "Measurement of Professional Attitudes," *Journal of Personnel Research*, V (1926), 131–41.

Freyd, Max. *Freyd's Occupational Interest Blank.* Chicago: C. H. Stoelting Co.

Garretson, O. K., and Symonds, P. M. *Interest Questionnaire for High School Students.* New York: Teachers College, Columbia University, 1930–31.

Hart, Hornell. *A Test of Social Attitudes and Interests.* University of Iowa Studies in Child Welfare, Vol. II, No. 4. Iowa City, Iowa: University of Iowa, 1923.

Minnesota Interest Analysis Test. In Donald G. Paterson and Richard M. Elliott, *Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press, 1930.

Stewart, Frances J., and Brainard, Paul P. *Specific Interest Inventory.* New York: Psychological Corporation, 1932.

Strong, Edward K., Jr. *Vocational Interest Blank.* Stanford University, California: Stanford University Press, 1930.

Composite and miscellaneous

Almack, John C. *Sense of Humor Test.* Cincinnati: C. A. Gregory Co.

Bernreuter, Robert G. *The Personality Inventory.* Stanford University, California: Stanford University Press, 1931.

Heidbreder, Edna. *Minnesota Personal Traits Rating Scales: Introversion-Extroversion and Inferiority Attitudes.* Chicago: C. H. Stoelting Co.

Humm, Doncaster G., and Wadsworth, Guy W., Jr. *The Humm-Wadsworth Temperament Scale.* Los Angeles, California: Doncaster G. Humm, 651 North Parkman Avenue, 1934.

Kent, G. H., and Rosanoff, A. J. *Kent-Rosanoff Free Association Test.* Chicago: C. H. Stoelting Co., 1910.

Laird, D. A. *Colgate Personal Inventory Rating Scales.* Hamilton, New York: Hamilton Republican, 1925.

Loofbourow, G. C., and Keys, Noel. *Personal Index.* Minneapolis: Educational Test Bureau, 1933.

Maller, J. B. *Character Sketches.* New York: Teachers College, Columbia University, 1932.

Neymann, C. A., and Kohlstedt, K. D. *Neymann and Kohlstedt Introversion-Extroversion Test.* Chicago: C. H. Stoelting Co., 1929.

Pressey, S. L., and Pressey, L. W. *Pressey X–O Tests for Investigating the Emotions.* Forms A (Adult) and B (Child). Chicago: C. H. Stoelting Co., 1920.

Rogers, C. R. *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age.* Teachers College Contributions to Education, No. 458. New York: Teachers College, Columbia University, 1931.

Rorschach, H. *Rorschach Ink Blot Test.* Chicago: C. H. Stoelting Co.

Willoughby, Raymond R. *Willoughby E M Scale.* Stanford University, California: Stanford University Press, 1931.

Chapter IX

TECHNIQUE AND THEORY OF MENTAL TESTS

I. Subject-Matter of Tests and Related Problems

THE principles of technique and theory which are discussed in the present chapter will be taken up in connection with the practical problems in which they arise. These practical problems are met with in the two situations of designing mental tests, on the one hand, and administering them, on the other. Most of the theoretical and technical questions concerning mental tests arise in either one or both of these two situations. Some of them also arise, it is true, when we attempt to interpret the results of mental tests. There is some overlapping between the questions which come up in the design and administration of tests and in their interpretation. In the present chapter we shall deal with these questions primarily from the point of view of design and administration. In a later chapter we shall approach the problems from the point of view of interpretation. In this later discussion it will be possible to assume familiarity with the treatment of the problem which we shall make in the present chapter.

The approach to the technical and theoretical problems from the point of view of design and administration does not mean that one may not be concerned with these problems except as he intends to design a mental test, or even to administer it. The subject is approached in this way primarily because it constitutes a convenient method of organizing the questions which will be taken up. In addition to this advantage, this mode of approach will be serviceable in case one wishes to design a test, or to examine a test from the point of view of its conformity to the technical requirements.

1. Selection of subject-matter

The first problem which we meet either in designing or judging a mental test is concerned with subject-matter. By subject-matter we mean the content of the test. We may think of the content from one of three points of view. We may either think of the material of which a test is composed, such, for example, as a list of words to which opposites are found, or a list of words which must be defined, or a list of arithmetic problems which must be solved, and so on. On the other hand, we may think of subject-matter from the point of view of mental process, or mental capacity, which the test is designed to measure. Thus we may think of a test as measuring memory or discrimination between the pitch of sound or the tones of color, or the ability to associate, or the ability to reason, etc. Or, in the third place, we may designate the test, not in terms of the objective material of which it is composed, or of the mental process which it is supposed to measure, but of the operations which this individual goes through in attempting to pass the test. This point of view may be regarded as intermediate between the other two. Whether we define the subject-matter in one or the other of these three ways, the problem is the same — What is the best subject-matter for a given purpose?

We have already met this question repeatedly in our historical account of the development of tests, and in our description of the various types of tests which are in common use. The function of our present discussion will be to bring together in an organized whole the problems which are related to the subject and which have been touched upon in a piecemeal way in the previous chapters. At the outset of our discussion of subject-matter we must draw a general distinction. This is the distinction between a test of a special mental process or of special capacity, on the one hand, and a test of general capacity, on the other. The

subject-matter of the test will be determined, in the first place, according as our purpose is to measure special capacity or general capacity. This is true, even though we should come to the final conclusion, as some do, that a test of general capacity is merely a collection of tests of special capacity. Even if this assumption is true, it is still necessary to determine what collection of special capacities is satisfactory as a measure of general intelligence or general capacity.

2. Subject-matter in tests of special capacity

It has already been pointed out that the early tests were designed chiefly to measure a variety of special capacities. During the development of tests the emphasis has shifted more and more to the measure of general capacity or general intelligence. In spite of this shift in emphasis, it is still necessary for certain purposes to measure special traits as distinguished from general ones. This is particularly true in tests for vocational selection and guidance. Not all vocational tests are of the specialized sort, but some of them clearly are. Perhaps the best examples of such tests are the Seashore Music Tests. Other examples may be found in the various monographs which describe the construction of tests to measure aptitude for particular jobs. It may be possible, also, to distinguish special abilities or disabilities in the school, and to measure them by means of tests. Some attempts have been made to do this as, for example, that by Bronner.[1] It is well known, for example, that certain children, though normal in intelligence, have great difficulty in learning to read. Many tests have been devised for the purpose of analyzing the ability to read and of diagnosing the deficiencies of pupils who encounter special difficulty in

[1] Augusta F. Bronner, *The Psychology of Special Abilities and Disabilities.* Boston: Little, Brown & Co., 1926.

learning to read. A prominent example of collections of such tests is that made by Marion Monroe.[1] Special tests may also be used in making a study of racial differences, or of differences between individuals in hereditary aptitudes, or in measuring the effect of environment and training.

How, then, must we proceed to the selection of the subject-matter for a specialized test? Obviously the first step is to locate the ability, and to attempt to define and analyze it. Sometimes, although the ability may rightly be described as specialized, it is by no means simple, or unanalyzable. Musical ability, for example, has been analyzed by Seashore into some thirty components. It includes pitch discrimination, discrimination of the intensity of sound, recognition of rhythm, recognition and memory of melody, discrimination between different degrees of harmony, the recognition of the relation between time intervals, motor dexterity in musical performance, control of the voice, musical appreciation, and many others. There is good evidence that not even these elementary capacities, or specialized capacities, are simple and unanalyzable. At least the capacities as we measure them cannot be regarded as ultimate units of mental ability. If we measure discrimination of intensity by various instruments, for example, we find a variation in the result. This means that discrimination of the loudness of one kind of sound is not exactly the same as discrimination of the loudness of another kind of sound. We may show from another point of view, also, that these tests, as they are ordinarily given, do not measure single unitary capacities. The ability to respond to any one of these sensory tests involves not only the capacity to discriminate the sensations themselves, but also the ability to pay attention and the willingness to pay attention.

[1] Marion Monroe, *Children Who Cannot Read.* Chicago: University of Chicago Press, 1932.

Probably we must regard the ability to pay attention as a somewhat generalized mental attitude. A specialized test, therefore, is to some degree, at least, a general test. When the psychologist confesses that he does not know exactly what it is that is being measured by the test which he uses, the statement does not mean that he has not made an effort to analyze the mental processes, but it means that he has found them to involve a complexity which he has not yet been able completely to resolve. It is the person who describes quite confidently the mental processes which are measured by his test who displays his ignorance of the subject. We cannot, of course, ultimately rest content with the failure to designate what it is that is measured by a test. Future research must be directed largely toward this problem.

In the meantime, we can meet our practical needs in two ways. In the first place, we can make a provisional analysis and give a provisional description of the mental process which we believe to be measured by the test. In the second place, we can take as the units which are to be measured the activities which are required in particular practical situations of life. Thus, if we cannot with clearness isolate and define memory, or if we find that memory is actually a composite of many simpler functions, we may, at least, measure the ease and rapidity with which a person learns poetry, or with which he memorizes telephone numbers, or the names of persons, and so on. In the army a group of psychologists wished to devise tests to measure aptitude for learning to fly. They took as the measure of flying ability, not an abstract measure of any sort, but the actual rapidity with which an individual learned when he undertook to master the aeroplane. They then selected tests to measure this ability by trying out a number which seemed likely to be successful. Then by correlating them with rapidity of learning they determined empirically which ones were successful.

We see that the attempt to design tests of specialized ability may proceed from two purposes. In the first place, the aim may be to make a scientific and accurate analysis of the components of mental capacity, or of the various specialized mental capacities. In the second place, the aim may be to measure the aptitude which is required to perform some particular activity in practical life. The theoretical or scientific problem is by far the more difficult, and we have made little progress toward its solution.

The second step, after we have defined the ability which is to be measured, is obviously to invent some means of measuring the ability in question. The method which has sometimes been used is that of analysis. The psychologist attempts to define to himself in psychological terms the nature of the ability, and then to assemble, or to invent, tests which may be assumed to measure the capacity which is thus analyzed. The problem is approached, in other words, from the *a-priori* point of view. On the other hand, the attempt may be made from a purely empirical point of view. The experimenter may try out one test after another without having any particular reason to expect that one test will be more successful than another. He then selects the test or tests which prove by experience, or empirically, to work, or to be successful. If several tests have been found by this procedure to be moderately successful, they may be combined into a team of tests.

The method which is most likely to be successful, and to reach the solution most quickly, is a combination of these two. No experimenter, in fact, ever proceeds in a purely random fashion. He makes a rough guess at the tests which he thinks will be successful and then proceeds to try them out. He sometimes wastes time, however, by not making as careful preliminary analysis as he might. This careful analysis may lead either to the trial of tests which otherwise

might not be thought of, or to the invention of the tests which may be more satisfactory than any that are in existence at the time.

The opposite error to that of failing to make a careful preliminary analysis is the one of resting content with analysis, assuming that a test will be successful without finding its correlation with some outside measure of achievement. Many examples of this error could be found in the earlier period of testing, though relatively infrequent at the present time. The correct procedure may be illustrated in the field of typewriting. If we wish to devise a test to measure aptitude for learning to typewrite, we may first, through whatever psychological insight we may have, assemble or devise a series of tests. Each one of these tests must then be put to a trial by giving it to a group, at the same time measuring the rapidity and ease with which they learn to use the typewriter, and then finding the correlation between their standing on the test and their score in learning.

3. Subject-matter of tests to measure group factors or primary abilities

The subject-matter of tests of special ability or of aptitude is selected, as has been described, by a common sense or subjective analysis of the ability to be tested and of the activities which represent the ability, or by a combination of this method and the empirical method of correlating tests with a practical activity which serves as a criterion. The subject-matter of tests of group factors or of primary abilities — the two may be regarded as synonymous — is selected by factor analysis. The abilities and the tests may resemble those which might be selected by common sense, for example, tests of memory, but the origin is different.

The process of factor analysis has already been described.

The method of selecting material involves first identifying the factors for which tests are to be designed. There are several statistical procedures for finding these factors, such as the two-factor analysis of Spearman, the bi-factor analysis of Holzinger, and the multiple-factor analysis of Thurstone. The factors are in the first instance assumed elements in the performances carried on in making the responses to various tests, which in combination make up the performances. Thus one test calls forth a performance which involves the recall of experiences plus verbal responses. Another involves making discrimination between musical tones plus motor responses, etc. Each of these elementary performances is supposed to be the expression of an ability. The performance on tests, therefore, is the expression of a combination of abilities or a combination of factors. The factors which are common to different tests are responsible for the correlation between these tests. Factor analysis consists in finding factors to account for the correlations.

In attempting to identify factors and to discover tests for measuring them the experimenter usually starts with a number of tests, each one of which he judges by common sense to depend on a certain ability. He then seeks by factor analysis to discover the smallest number of factors which will account for the correlations between the tests. Next, he determines the "loadings" of each test with the various factors, that is, the proportions in which each factor is represented in each test. He selects tests which have high loadings with particular factors, or high correlations with these factors, and have loadings with others, and examines them to try to give a psychological meaning to the factor. That is, he applies common sense or psychological insight again to form an idea of the nature of the ability represented by the factor. At this point he usually gives it a name in accord with ordinary psychological terminology. Finally, he

tries to revise the test so it will be a purer measure of the factor in question, and checks up on his procedure by making a factor analysis of the revised set of tests. This last step has not been taken with sufficient completeness to furnish an adequate set of factors and tests to measure them.

The two groups of factor analysts differ in their analysis of abilities in one essential point. Those of the Spearman school distinguish one general factor, *g*, a number of group factors, and factors which are specific to each test or performance, *s*. Other analysts, like Thurstone and Kelley, account for all abilities in terms of abilities which are not general or universal. Thurstone calls these factors primary abilities.

The two groups agree in positing a limited number of

TABLE XI. FACTORS IN ABILITY SUGGESTED BY HOLZINGER, KELLEY, AND THURSTONE

Holzinger	Kelley	Thurstone
1. General		
2. Mathematical-mechanical	Facility with quantitative concepts	Number facility
3. Verbality	Facility with verbal material	Word fluency
4. Spatial factor	Facility in mental manipulation of spatial relationships	Visualization of space
5. Memory	Memory facility	Memory for words, names, and numbers
6. Mental speed	Speed in mental processes	Perceptual speed
7. Verbality		Verbal reasoning
8. Deduction		Induction
9. Motor speed		

abilities which run through a series of tests or performances, but not all. These will be mentioned here. The testing of *g* will be discussed under testing of general intelligence.

The two groups and the various investigators agree fairly well, also, as to what the group factors or primary abilities are though they do not believe that they have yet identified all of them. It will be seen that there are five factors in which the three substantially agree. Induction is regarded by Thurstone as corresponding to Spearman's *g*.

It is worthy of note that the factors identified in the studies to date are very similar to those which have long been identified by common sense. Is this a confirmation both of common sense and of factor analysis, or is it an evidence that factor analysts have been influenced by their ordinary psychological concepts in the search for factors? Further discussion of the nature of ability and its components will be given in a later chapter.

4. Selection of subject-matter for tests of general intellectual capacity — The existence of general intelligence

Before discussing the kinds of tests which are adapted to measure general capacity, and before indicating what the tests of general capacity should be like, it is pertinent to raise the prior question whether general intellectual capacity exists. The existence of general intellectual capacity is not universally accepted, and there is considerable debate concerning its nature among those who do accept its existence.

We may distinguish three general conceptions about the existence of general intelligence. The first is that there is such a thing as intellectual capacity which enters more or less into the performance of all intellectual work, or which constitutes a factor in every type of intellectual reaction to the world about us. This general factor is the same whatever practical situation it appears in, or whatever other

intellectual factor it is associated with. While not the sole factor in intellectual achievement, it is the most important one.

A second view is that ability is made up of a limited number of broad factors, which enter into a number of operations and therefore affect performance on a number of tests, but not all. This is the view which is expressed in the statement that total ability is made up of a few primary abilities.

According to the third view each ability, represented by performance on a test, is composed of an assortment of a large number of factors. Each factor may enter into many performances but the assortment will be somewhat different from performance to performance.

The design of a test of intelligence will doubtless be affected somewhat by the concept of ability which the designer entertains. However, it would not be correct to say that the tests have been worked out so as to exemplify one or the other of these concepts. Spearman, in fact, criticizes them because they are a heterogeneous collection of tests assembled empirically rather than with a clear concept of *g* and an attempt to design tests to measure. Thurstone advocates the substitution of tests of primary abilities for the traditional intelligence test. No proponent of the third view, so far as the writer is aware, has proposed a plan for designing tests in accordance with it. In the meantime, a review of the actual development of so-called intelligence tests may be illuminating as a basis for further discussion.

The earlier students of mental tests who wished to measure general intellectual capacity sought for some single test that would measure a particular capacity. That is, they identified general intelligence with one of the particular mental functions. There are several examples of this procedure. Spearman, in his early studies of correlation, found a certain degree of correlation between various tests

of sensory discrimination. He therefore concluded on this meager evidence that general intelligence consists of fineness of discrimination. He described the difference between high intelligence and low intelligence, figuratively, by comparing the former with highly tempered steel and the latter with iron. Ebbinghaus, it will be remembered, sought a measure of general intellectual capacity in the combining or associating process, and devised his completion test as a means of measuring this process. Binet, in his earlier experiments, chose attention as probably the most essential aspect of intelligence, and devised a series of tests to measure it.

As the experience with mental tests accumulated, it became more and more difficult to identify general intelligence with any one particular mental capacity. It was found that a number of tests proved to be successful, as measured by their correlation with criteria or other measures of intelligence. It was found, furthermore, that a combined score of a series of tests usually gave a higher correlation with the criteria than did the score from a single test. This has led to an attempt on the part of those who still define general intelligence as a universal capacity to describe it in more general terms. Possibly the first to formulate a clear definition of intelligence from this point of view was Stern. Stern defined intelligence as the general mental adaptability to new problems and conditions of life. At about the same time Burt, on the basis of his studies of correlation, defined the central factor as "the power of readjustment to relatively novel situations by organizing new psycho-physical coördinations." Similarly Binet, as quoted by Terman, describes intelligence as "(1) the tendency of thought to take and maintain a definite direction, (2) the capacity to make adaptations for the purpose of attaining the desired end, and (3) the power of self-criticism." Again, Colvin described intelligence as capacity to learn. These definitions agree in

a general way with the description of intelligence by James, from the point of view not of mental tests, but of a description of the mental processes and their development. James described intelligence, in contrast to instinct and habit, as adaptation to novel conditions by variation of behavior.[1]

It is difficult, however, to apply such descriptions as these to all of the parts of the prevailing intelligence tests. If one will examine the Binet scales or the various point scales one will see that they call for a variety of mental operations and that some of them at least do not call for the adjustment to novel situations or for new learning. Some, to be sure, do, as for example, the problem or puzzle type. Among the widely used language tests the analogies and completion tests are of this type. On the other hand, the vocabulary and opposites tests depend more largely on previous learning. This is also true of other tests in the various scales. Examples from the Binet are naming objects and counting.

Although it is not possible to bring all the accepted tests strictly under the category of adjustment to a novel situation, it is true that they do, in the main, present situations which are novel for the individuals at the intellectual level for which the particular tests are designed. Excluded are the tests which call for merely routine or habitual response and, for the most part, those of simple sensory discrimination and motor response. This justifies the statement that, while the tests call for some diversity of particular operations, they do emphasize doing something novel and, at the more advanced levels especially, utilizing ideas and carrying on abstract thinking.

The foregoing observation on the descriptive character of the intelligence tests which have come to be accepted leads to the notion that there is some sort of unity among the tests in spite of the diversity in their content. This concep-

[1] The definition of intelligence will be discussed more fully in Chapter XVI.

tion has been supported by the results of correlation between tests and has also led to the use of correlation between tests as a method of selecting tests for intelligence scales.

The theory that there exists some sort of general factor would seem to imply that those tests which have the highest intercorrelation would be the best measures of intelligence, and that tests of intelligence should be selected partly upon this ground. In the early stages of the work with the army tests a contrary theory was adopted, and for a time was followed. This theory is expressed in the following words: [1]

> The general principle is that the lower any particular test correlates with them, the greater weight it should have in the composite. For in the proportion that two tests intercorrelate closely, they are repetitive — i.e., are measures of the same fact — and a high weight to each of them will mean an undue weighting of the same fact. The lower the correlation of this fact with the fact to be prophesied, the more excessive would the weighting be.

As a result of their experiments in the design of the army test, however, the authors have the following to say:

> A test which will not correlate thoroughly well with the total score of a good battery of tests is *ipso facto* under grave suspicion; there is little likelihood that it will consistently correlate well with any other proved measure of intelligence (p. 338).

We shall see in a moment from the statistical evidence that this statement is correct.

In weighing this issue, we must distinguish two points of view, the mathematical and the psychological. From the mathematical point of view it may be proved that the guiding principle which was adopted at the beginning of the army testing work is correct. That is, if each one of a battery of tests correlates to a certain degree with a criterion, then the lower the intercorrelation between the tests, the higher will be the correlation of the composite score of all the

[1] Robert M. Yerkes (Editor), *Psychological Examining in the United States Army*, p. 316.

tests with the criterion.[1] This is a purely mathematical fact which has nothing whatever to do with the psychological make-up of intelligence. The psychological question, however, is an entirely different one. It is this: Are there tests, as a matter of fact, which have a low intercorrelation and which correlate to a high degree with a criterion; in other words, can tests be found which meet this mathematical desideratum? If they cannot be found, then it will be necessary to sacrifice either the requirements of a high correlation with the criterion, or of a low intercorrelation.

The psychological fact is clear, and can be demonstrated from the army test results themselves. It is displayed in Table XII, which has been put together from the various parts of the army test report. It contains, on the one hand, the intercorrelation of the various individual tests of the Army Scale A, which was preliminary to Scale Alpha, and also the correlation between these various tests and criteria. The references to the sources of these figures in the army report are given in the table. Following the presentation of the various correlations with criteria, and of the intercorrelations of the tests, are given the rank orders of these sets of correlations. The rank order of the correlations with criteria and the rank order of intercorrelations are each combined into a composite rank order. The essential comparison is between the composite rank orders. It is very evident from an inspection of these data that a test which has a high correlation with criteria also has high intercorrelations, and a test which has a low correlation with criteria has low intercorrelations. The two composite rank orders are almost identical.

It would appear from these facts, then, that for the pur-

[1] This is a deduction from Spearman's formula given in Charles Spearman, "Correlations of Sums and Differences," *British Journal of Psychology*, V (March, 1913), 417–26.

Table XII. The Relation between the Correlation of Tests with Criteria and their Intercorrelation in the Case of Army Test A

	Page in Army Report	Oral directions	Memory span	Disarranged Sentences	Arithmetic	Information	Opposites	Pract. Judgm.	Number Completion	Analogies	Number Comparison
		1	2	3	4	5	6	7	8	9	10
CORRELATIONS WITH CRITERIA											
1. Officers' rating of 313 National Guard	315	47[1]	34	48	46	54	51	41	42	36	47
2. Officers' rating of 338 Men	331	41	36	30	46	45	50	39	33	43	39
3. Mental age, average of 8 groups	332	47	36	49	59	53	67	54	43	59	36
4. Trabue B and C, 287 pupils	337	60	39	55	65	65	58	66	57	66	50
5. Grade location	337	49	40	56	67	70	66	58	61	76	55
INTERCORRELATIONS											
1. Average intercorrel. 313 National Guard	316	58	47	54	61	64	61	56	49	54	52
2. Average intercorrel. 895 Engineers	539	62	52	57	66	66	67	60	55	64	56
RANK ORDER OF TESTS IN CORRELATIONS WITH CRITERIA											
1.		4.5	10	3	6	1	2	8	7	9	4.5
2.		5	8	10	2	3	1	6.5	9	4	6.5
3.		7	9.5	6	2.5	5	1	4	8	2.5	9.5
4.		5	10	8	3.5	3.5	6	1.5	7	1.5	9
5.		9	10	7	3	2	4	6	5	1	8
Total		30.5	47.5	35	17.0	14.5	14	26.0	36	18	37.5
Composite Rank		6	10	7	3	2	1	5	8	4	9
RANK ORDER OF TESTS IN INTERCORRELATION											
		4	10	6.5	2.5	1	2.5	5	9	6.5	8
		5	10	7	2.5	25	1	6	9	4	8
Total		9	20	13.5	5	3.5	3.5	11	18	10.5	16
Composite rank		4	10	7	3	1.5	1.5	6	9	5	8

[1] Decimal points are omitted from these coefficients.

pose of selecting subject-matter for an intelligence test the intercorrelation between the tests results in practically the same selection as does the correlation between the tests and criteria. Furthermore, it is psychologically impossible to select tests which have a high correlation with criteria and yet which have a low intercorrelation. This being the case, and since it is obvious that one must select tests which have a high correlation with criteria, we must abandon the demand, which rests upon mathematical considerations, that the test shall have a low intercorrelation.

The facts which have just been referred to, in addition to the hierarchy of test coefficients, are regarded by some psychologists as supporting the hypothesis of a general factor in intellectual ability. Perhaps we can regard intelligence as a unitary, though complex, mental trait, or characteristic. Variations in achievement or productiveness in various human activities may be ascribed in part to differences in intelligence, and in part to differences in other mental traits than intelligence. The practice of test designers in selecting material seems to reflect a conception of intelligence involving unity in diversity.

It might perhaps be concluded from a logical application of the principle of the hierarchy of abilities that our best procedure would be to find some one test which has the highest intercorrelation and the highest correlation with criteria, and rely upon this alone. Experience has shown, however, that groups of tests give better measures than any single test. How is this to be accounted for?

The fact that no test is a perfect measure, and that groups of tests are better than single tests, may be accounted for on the ground that every test involves certain irrelevant factors as well as the central factor which we are attempting to measure. To put it in another way, every test measures other capacities in addition to intelligence. Intelligence always operates upon material. We think in terms of our

experience, and not in terms of purely abstract ideas unrelated to the world of experience. Our ability to deal with any materials of thought, then, will depend, not simply upon thinking ability in the abstract, but upon our familiarity with and our ability to deal with particular materials with which the thinking is to be carried on. For example, thought may be carried on in terms of language. It may be carried on again in terms of mathematical symbols or in terms of mechanical relationships. The thought activity which is carried on with these different types of materials, or with these different modes of expression, may be the same in its general character. The skill and ease with which an individual may carry on a train of thought, however, will depend, in part, upon the nature of the materials and his adaptation to them. Take a concrete example. A lawyer is able to reason in terms of legal facts and principles; a physician is able to reason in terms of medical facts and medical laws; the engineer can reason in terms of the physical relationships of material things and their laws. The thought process may be abstractly the same in all these cases, but the mental process is colored by the material of thought as well as by the form of thought.

If the distinction between the material and form of thought which has just been drawn is correct, we have a justification for the use of a variety of tests. The need for this variety is due, not so much to the fact that different mental processes are measured by them, as that each of them measures the mental activity only as it appears in certain concrete operations of thought, and that these different concrete operations are conditioned partly by their material embodiment, as well as by their form.

If we carry this line of thought a step farther, we are led to raise the general question of the validity of our intelligence tests. The question is whether or not all of our tests

are not limited by the fact that they deal with a certain restricted range of thought material. Within this range of material, they have demonstrated validity. Their range has been limited chiefly, however, to the realm of school activity. Would they have the same validity if they were applied in a variety of other situations outside the school? Is school intelligence the same as life intelligence?

We must undoubtedly answer this question by saying that the measure of achievement in the school is not identical with the measure of achievement in other situations outside the school. To say that it is not identical, however, does not mean that there is no relationship. The degree of closeness of this relationship we shall have to consider in dealing with the practical application of tests. For the present we may say that the relation of the test scores to school success is closer than its relation to success in life outside. This is due to the fact that the materials of the test are largely materials of school work. How, then, shall we interpret this lack of identity? Shall we say that there is a school intelligence and that there is a life intelligence, and perhaps that there are various kinds of life intelligences, or shall we say that intelligence is measured to some degree by tests which involve typical school activities, but that the measure is limited by the fact that the material in which the tests are represented is of a somewhat specialized nature?

Possibly the question is one of definition. To the writer, however, it seems to be the simpler way of expressing the facts to say that the intellectual activity required in the various situations of life is similar in character to that required in school, but that one's achievement in any particular case is conditioned by materials of thought and his familiarity with them, as well as by the form of thought.

Whichever interpretation is the correct one, the fact remains that our so-called intelligence tests have limitations.

This limitation is sometimes disregarded with serious consequences. For example, when the scores made by the adults in the army on the Stanford-Binet test are given the same significance as the scores made by school children on that test, a serious error is made. The same error is made when the scores of immigrants who have lived in the United States for fifteen or twenty years are compared directly with the scores of immigrants who have been here but a few years, and are treated as having the same significance.[1] The danger of misinterpretation of the test scores has led some to hold that the term intelligence test is an unfortunate one, and that we ought to call the tests academic ability tests, or something of the sort. If we retain the name, as we probably shall on account of its wide acceptation, we shall do so only with the distinct understanding that their application is affected by the fact that they deal primarily with school material.

We have seen that the standing in intelligence tests is determined in part, not only by the individual's native intellectual capacity, but by the nature of his past experience. The aim of intelligence tests is, so far as possible, to so choose the materials of which they are composed that the effect of differences in experience will be reduced to a minimum, and this aim has in a measure been attained. No one would claim, however, that the attempt has been completely successful. That specific teaching of subject-matter similar to that which appears in an intelligence test produces a marked gain in the scores in the test is shown in an experiment by Bishop.[2] Groups of high-school pupils were given special drill in handling problems similar to those in the Otis Group Intelligence Scale, and their gains were compared with paired groups who were given the test twice

[1] For fuller discussion of this problem see Chapter XV.

[2] Omen Bishop, "What Is Measured by Intelligence Tests?" *Journal of Educational Research*, IX (January, 1924), 29–38.

without such drill. The trained groups made gains from two to seven times as large as the check groups. This training came very near to being direct coaching, and it is not likely that ordinary differences in schooling would produce such large differences, but it is clear that intelligence tests are by no means independent of schooling.

We now approach another question related to the first one, but not identical with it. This question is whether general intellectual capacity is to any degree specialized on account of differences in aptitude for dealing with different types of problems. Part of the apparent specialization of general intellectual capacity may be ascribed to the combination of intelligence and non-intellectual traits. For example, two persons of equal intellectual capacity may make a very different impression upon their associates in personal intercourse, because of differences in their personalities and in their reactions to social situations. The one, for example, may be timid, and may become confused when he attempts to express his thoughts to another individual or to an assembly. The other person may be actually stimulated by the presence of others so that he thinks more clearly than when alone. The one may have presence of mind in an emergency, while the other loses his head. The one may do better under the stimulus of competition, the other when he is impelled by intrinsic motives alone.

Another possible explanation of the individual's variation in the performance of various intellectual tasks is the possible effect of variation in what we have designated as specialized factors, in association with general intellectual capacity. We have regarded manual dexterity, for example, as a capacity which is largely specialized. It differs among individuals largely independently of general intelligence. However, certain tasks requiring general intelligence for their performance require also a certain degree of manual

dexterity, or may be performed better by an individual with a high degree of dexterity than by one with a low degree of dexterity. Other tasks, on the other hand, demand language ability, and this may be regarded, to some degree at least, as specialized. Other intellectual operations, such as those of mathematics, require the manipulation of abstract symbols. The facility in the use of symbols may possibly be somewhat specialized.

Whatever may be the explanation of the fact, the basic fact of practical importance is this: While we aim to measure general intelligence, and while we believe we can with some degree of precision measure it by means of our tests, we are never able to use perfectly general or abstract material. Our material is always particularized, and the test is therefore to some degree rendered a specialized test. It is specialized both from the point of view of previous experience of the individual being tested and from the point of view of special aptitudes. This particularization, or specialization of the test, is not sufficient to invalidate it, but it is sufficient to make it necessary to take account of certain limitations in the interpretation of the scores. To express these limitations in a sentence, we may say that a general intelligence score is always to be regarded as an approximation and not as a perfect measure of the intelligence, either of the individual or of a group.

This general conception of intelligence and of the kind of tests which are useful in measuring it has been attacked on two sides, from the points of view of the first two conceptions of intelligence mentioned at the beginning of the chapter. The advocates of both theories regard it as vague and empirical and out of harmony with the findings of factor analysis. Spearman would substitute a more clearly defined factor, *g*, for what he regards as the vague and loose conception represented in the current intelligence tests.

He has undertaken to define *g* psychologically as well as to devise tests to measure it. It consists of two processes, the eduction of relations and the eduction of correlates. In addition to this, he posits group factors which are to be tested separately. Thurstone and Kelley, on the other hand, also on the basis of factor analysis, believe that what we call intelligence is really a composite of coördinate factors, which they call primary abilities or mental traits. According to their view, each of the abilities should be tested separately. A composite score would represent merely a mathematical average and not any general or central ability.

It is necessary at this point only to indicate the differences in these views and their obvious bearing on the design of tests. They will be discussed further in the chapter on the nature of ability.

Chapter X

TECHNIQUE AND THEORY OF MENTAL TESTS

II. Problems Relating to the Selection and Organization of the Items of a Test

1. Principles concerned with selecting the items of tests

After it has been determined what kind of test material shall be used in a test, the next task is to find particular material of the kind which has been determined upon. We need not consider here in detail the way in which the maker of a test goes about it to find appropriate material. He sometimes makes use of test material which has already been devised by somebody else. In some cases this material is modified, and in other cases new material of a similar sort is discovered or invented. In some cases a radically new type of material is devised. We may assume that, by any means at his disposal, the person who is to construct a test has made a collection of a large number of items of the sort which he wishes to use. We are concerned now with the principles which should guide him in the selection and arrangement of this material.

The first fundamental principle is concerned with the difficulty of the items which are to be included in the test. The first requirement is that the items be of a suitable general level of difficulty. The test must be difficult enough and not too difficult for the individuals for whom it is designed. If it is for primary children, it must be of one level of difficulty; if it is for adults, it must be of a very different level. A test does not usually, as we have seen, represent a single dead level of difficulty. The items cover a certain range.

Thus a primary test may be suitable for administration to children of from six to ten years of age. A few tests, as we have seen, have such a wide range of difficulty that they are designed to be applied to all ages of individuals from the third year up. This means that some items are very easy, and some are very hard. The test as a whole does not represent any one level, although individual parts of the test do.

With respect to the relative difficulty of the various items or parts of a test a choice confronts us. We may seek either to make all of the items of the same difficulty, or we may select items which are graded in difficulty throughout either a narrow or a wide range. The maker of a test, in short, has to choose whether it will be of uniform difficulty or of graded difficulty.

What are the considerations which will determine this choice? We may assume that the test is to be used to differentiate between the abilities of a group of individuals. This assumption means that, after the test has been given, the individuals may be arranged in a series or in a distribution table according to the scores which they make on the test.

Consider first what is the basis of this ranking of individuals in case the items of the test are all of the same difficulty. In order that all the individuals of a group may score, the test must be easy enough so that everybody can pass the various items. If an individual can pass one of the items he can pass all of them, or at least most of them. If the test is so difficult that some of the individuals cannot pass it, or cannot do any of the items, it will not be possible to arrange all of the persons in a series. If the test is easy enough for everybody to pass it, the differences in score will be determined chiefly by the speed with which the different individuals work. The speed of working may be deter-

mined in part by the relation between the difficulty of the test and the individual's capacity, but the test will be chiefly a measure of speed. It seems best that the items of a test be graded in difficulty to include some which nearly all the individuals will pass and some which few will pass. This will best differentiate between the ability of individuals throughout the range. Thus, the average difficulty of the items will about correspond to the average ability of the group to be tested, that is, the median item will be passed by about 50 per cent of the group.

If the test items are graded in difficulty, it is at least conceivable that they may measure, not primarily speed, but the limits of capacity. Consider, for example, a number completion test. It is possible to construct number completion items which increase in difficulty to a point at which most persons would fail. In fact, it is conceivable that the ablest mathematical thinker in the world could construct a test on which everybody else would fail at some point. A test like this would measure the limits of capacity, or, to use a term contrasted with speed, power.

It would perhaps be impossible to construct a pure speed test or a pure power test. The differences in speed are caused partly by differences in power. Most tests do not range in difficulty beyond the limits of capacity of at least the abler individuals who are tested. If they were given time enough, they might pass the entire test. Tests graded in difficulty are usually given with a time limit. While tests usually measure neither speed nor power exclusively, they may be predominantly the one or the other. They may be easy and given with a time limit so set that nobody can finish within the limitation, or they may be steeply graded in difficulty and given with a liberal time allowance.

There has been some inquiry as to whether intelligence tests measure chiefly speed or power. The earliest attempt to study this question was made with the Army Scale

Alpha.[1] The study in question consisted in determining the effect of doubling the time of the test. This effect was studied in two ways, first, by finding the correlation between the scores on single time and on double time, and second, by finding the percentage of individuals at various levels who improved their score when the time was increased.

With reference to the correlation between the tests on single time and on double time, the authors of the report assumed that if the test is a speed test, the correlation would be high. The correlation could be low only in case the test is a power test. To quote the report:

> A change of order would occur only if the test were of the type in which time was relatively unimportant — so-called "power" test. Here it might happen that quick individuals scoring high would reach the limit of their abilities and fail to profit by additional time, whereas slow, capable persons would plod unerringly on in the extended period and outdistance in the end their more speedy rivals.

It was found, in fact, that the correlation between single and double time was very high, being .965. On the assumption just mentioned, then, we should conclude that, in so far as the correlation fact is concerned, the Army Alpha test is largely a speed test.

It is interesting to note that Brigham has drawn the opposite conclusion from the same fact.[2] Brigham takes up the criticism which has been made of the army tests that they are chiefly speed tests, in that they penalize slow, but ac-

[1] Robert M. Yerkes (Editor), *Psychological Examining in the United States Army*, Chapter IX, Part II. See also G. M. Ruch and Wilhelmine Koerth, "'Power' *vs.* 'Speed' in Army Alpha," *Journal of Educational Psychology*, XIV (April, 1923), 193–208; and Giles Murrel Ruch, "The Speed Factor in Mental Measurements," *Journal of Educational Research*, IX (January, 1924), 39–45.

[2] Carl C. Brigham, *A Study of American Intelligence.* Princeton, New Jersey: Princeton University Press, 1923.

curate individuals. He refers to the high correlation between the scores on single time and double time, and concludes, "At least in our consideration of the army test results, we may definitely discard the opinion that we are testing speed rather than intelligence" (page 12). Brigham's argument seems to be that because, when we extend the time, the individual who made a high score increases his score, as well as the one who made a low score originally, the person making the low score was not penalized by the time limit. But this is not the issue. The question is not whether the time limit was too short to allow the slow individual to make a score. Suppose that two men were running a race. This would certainly be a speed test. Suppose the race were not long enough to test endurance and were long enough to be unaffected by the differences in reaction time at the start. Suppose that we scored them in distance at the end of ten seconds, and one had run ninety yards and the other one hundred. Suppose then, we scored them again at the end of twenty seconds. If they maintained the same speed, the first one would have run something like two hundred yards, and the second something like one hundred and eighty yards. Doubling the time would not affect their relative scores.

If the test is a speed test it seems pretty clear that doubling the time will result in a very high correlation of scores, assuming that no considerable number of individuals stop before time is called. It seems doubtful, however, that we can reverse the argument and say that because there is a high correlation the test is necessarily a speed test. This depends upon yet another fact, the correlation between speed and power. If there is a high correlation between speed and power, so that the person who works rapidly also has high degree of power, then doubling the time on either a power test or a speed test would give a high correlation

between the scores on single time and on double time. This is because a measure of speed is also a measure of power. If, on the other hand, there is a low or zero correlation between speed and power, then we should expect that doubling the time on a power test would give at least a much lower correlation between single time and double time scores. It is probable that the correlation between speed and power is positive, but moderate. If this is the case, and our reasoning is correct, a very high correlation between single and double time indicates that the score on the test depends largely upon speed of performance.

The writer attempted to test these assumptions by a simple experiment. The plan was to construct a predominantly speed test and a predominantly power test of the same kind of material, namely, number completion examples. In the one the examples were of approximately equal difficulty; in the other they increased in difficulty. The tests were given to a university class and 46 complete records were obtained. The correlation was found between the scores on single time and double time for both tests, and also between the scores on the two tests. They were as follows:

Correlation between single and double time, speed test	$.87 \pm .03$
Correlation between single and double time, power test	$.78 \pm .04$
Correlation between speed and power test	$.63 \pm .06$

The comparative lowness of the correlation between the speed and power tests shows that there is a distinction between speed and power, and that they are at least not perfectly correlated. The lower correlation between single and double time in the power test than in the speed test shows that if a test gives a very high correlation between single and double time the indications are that it is a speed test. Hence the Army Alpha Test appears to be largely a speed test, as the authors of the army report maintain. The data

suggest, finally, that if the purpose is to measure power, it is necessary to allow ample time for its performance.

This conclusion is confirmed by an investigation of the matter by Odoroff.[1] This author uses a new technique to rid the correlation between single and double time of the disturbing factor of the size of the score on single time, which is common to both single and double time scores. He expresses the score on extended time as a deviation from the probable score. He then finds that when the mean score on a test is high, that is, the test is easy and is largely a speed test, the correlation between original and extended time is high and vice versa. A hard test differentiates on the basis of power and an easy one on the basis of speed.

The army psychologists attacked the question in a second way. They calculated the percentage of individuals who gained in score on each test at the various levels. That is, they found the percentage of individuals who made a score of 1 on single time and who increased their score on double time, similarly for those who made a score of 2, 3, 4 and so on, up to the highest score made. The hypothesis was that if the tests were speed tests, the large majority of persons would gain on double time at all levels, but if they were power tests, a small percentage would gain on double time. What they found was that a larger percentage of those making high scores than of those making low scores on single time gained with double time. They concluded, therefore, that for the low grade individuals the tests were largely power tests and for the high grade individuals largely speed tests. This seems to be a reasonable conclusion. They concluded, also, that on the whole the power factor is not so important as the speed factor even at the lower level.

[1] M. E. Odoroff, "A Correlational Method Applicable to the Study of the Time Factor in Intelligence Tests," *Journal of Educational Psychology*, XXVI (April, 1935), 307–11.

Whatever may be the fact concerning the army test, we have here an issue which is important. It is probable that for some purposes, and in some situations, speed of performance is the important qualification. In other cases, however, the speed is relatively unimportant, and power, or the capacity to succeed in a task which is too difficult for other individuals, is the important characteristic. In typewriting and stenography, for example, speed is the important factor. Speed, of course, is not here contrasted with accuracy, and accuracy is not to be identified with power. What has been said about speed is true of effective speed, or speed combined with accuracy. In invention and creative scientific work, on the other hand, power is relatively much more important than speed. It is of little moment whether Newton, and Darwin, and Tesla, and Edison, and Einstein developed their theories or perfected their inventions in one year, or ten years, or twenty years. The important thing is that they were able to perform intellectually far above the average individual. Their work possesses an importance measured in terms of quality, rather than of quantity.

The question of the relation between speed and power or altitude in a given performance is sometimes confused with the question whether there is a general factor of speed in all kinds of activities and whether this coincides with a general factor of intellectual performance. Two early studies by Travis [1] and by Peak and Boring [2] seemed to show that this was the case. This was a startling revelation, if true, because it was contrary to the general trend of findings on the correlation of speed of simple activities and intelligence. But Peak and Boring experimented with only five subjects

[1] Lee Edward Travis and Theodore A. Hunter, "Relation between Intelligence and Reflex Conduction Rate," *Journal of Experimental Psychology*, XI (October, 1928), 342–54.

[2] Helen Peak and Edwin G. Boring, "Factor of Speed in Intelligence," *Journal of Experimental Psychology*, IX (April, 1926), 71–94.

and Travis [1] completely reversed his finding in a later study! We must conclude, therefore, that speed in simple activities is not a good indicator of intelligence, as well as that speed in a given intellectual activity is not identical with power or altitude in that same performance.

The conclusion, then, is that not all intellectual performances can be measured in the same dimension. Some need to be measured, perhaps, largely in the dimension of speed, some in the dimension of quality or power, and some in a combination of both. Most of our tests measure a little of both, with probably the greatest emphasis on speed. This may be the most serviceable, in so far as a single measure is concerned. It seems likely, however, that it would be desirable to secure an analysis of the individual's ability by a test, part of which should depend largely upon speed of performance, and part upon power. In this way we could establish, not only the individual's general score, but also his capacity in each of these two characteristics separately.

The discussion of power tests and speed tests, it will be recalled, was introduced in the consideration of the two ways of organizing the items of a test, namely in a series of equal difficulty or in a series of graded difficulty. We may comment further on the procedure which is to be followed if the items are to be graded in difficulty. When test items are to be arranged in order of difficulty it is usual to take as the measure of difficulty the percentage of individuals who pass the test or who fail upon it. Thus a hard item, of course, is one on which a large percentage fail.

The attempt is usually made to select items which shall be graded in equal steps or intervals in difficulty. In determin-

[1] Lee Edward Travis and Clarence W. Young, "The Relations of Electromyographically Measured Reflex Times in the Patellar and Achilles Reflexes to Certain Physical Measurements and to Intelligence," *Journal of General Psychology*, III (July, 1930), 374–400.

ing the steps in difficulty the assumption is made that abilities are distributed according to the normal probability curve. This means that at the upper and lower extremes of the scale a given step or unit will be represented by comparatively few individuals, whereas in the middle part of the scale of difficulty a unit of ability will be represented by a large number of persons. Since the persons who possess abilities which are measured by the various units of the scale are not equal in number, it is not correct to determine the units of the scale directly by equal numbers of percentages of passing or failing. It is necessary to find the percentages failing which correspond to equal steps on the scale and then to find items which are failed by these percentages. The scores representing equal steps of difficulty are called percentile scores. A table for transmitting percentage of failing into percentile scores may be found in Rugg's *Statistical Methods Applied to Education*, p. 392.[1]

The other method of scaling the difficulty of parts of a test is the mental-age method. We have already seen how this method is applied in the development of age scales. In the application of this method we assume that a test which is passed by a given percentage of the older children is more difficult than one which is passed by the same percentage of younger children. Tests may therefore be arranged in the order of difficulty by taking those which are passed by a given percentage of children of succeeding ages. We cannot assume, however, that the differences in difficulty between the tests, when they are selected by this method, are equal. This would be assuming that the growth in mental capacity from age to age is uniform. Whether this is true or not is a matter which cannot be assumed, but must be determined by other methods of investigation, or on the basis of tests

[1] Harold O. Rugg, *Statistical Methods Applied to Education.* Boston: Houghton Mifflin Co., 1917.

which are standardized in another fashion. An age scale, since it is standardized in terms of age growth, can give us no information concerning the form of the curve of mental development.

2. The number of items in a test, or the length of the test

It was remarked in Chapter VI that there is reason to suppose, theoretically, that a longer test is more reliable than a shorter one. Certain authors of tests have applied this principle to test construction by employing considerably more material than is usual, and by correspondingly lengthening the time required to take the test. Thus, the Thorndike Intelligence Examination for High School Graduates occupies three hours. The question now before us is whether it can be determined just what the relation is between length and reliability, and whether from this it can be determined how long a test should be.

To determine the relation between length and reliability a formula has been devised by Spearman and applied experimentally by Holzinger.[1] If we have a series of similar tests of equal length and reliability, and we know the reliability of each of the tests or components, we can predict from this formula what the reliability of a composite of any number of the components should be. Of course, the components of a test are not entirely similar nor of equal length or reliability, and therefore the formula cannot be applied rigidly, but Holzinger has shown that actual test scores approach this law as the characteristics of the tests approach the assumed characteristics. In the case of the Otis Self-Administering Test, which is typical of the tests we are here interested in, the actual increase in reliability, in comparison with the

[1] Karl J. Holzinger and Blythe Clayton, "Further Experiments in the Application of Spearman's Prophecy Formula," *Journal of Educational Psychology*, XVI (May, 1925), 289–99.

theoretical increase according to Spearman's law, is shown in Fig. 13 (p. 293, *op. cit.*).

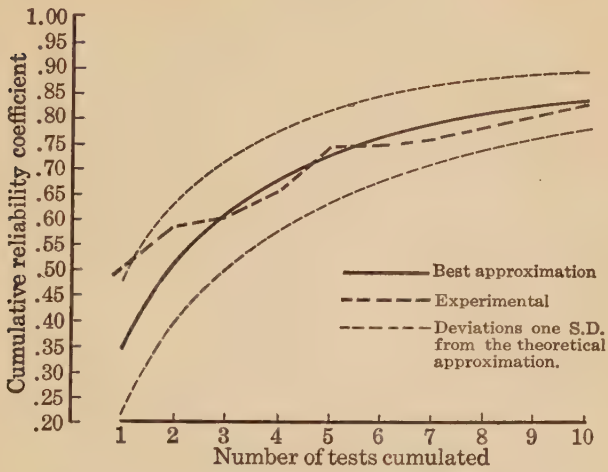


FIG. 13. EXPERIMENTAL TREND COMPARED WITH BEST APPROXIMATION TO SPEARMAN'S PROPHECY LAW

Otis Intelligence Test data. (From Holzinger and Clayton.)

We may conclude tentatively from the facts before us, that for tests composed of the material commonly used in our intelligence tests, and for components which occupy one and one half minutes working time (the divisions used in the above experiment), an increase in the number of components up to five causes a marked increase in reliability, and that there is a continued slight increase in reliability up to at least ten units.

3. The form of organization of the items of a group language test

It was remarked in discussing the development of group tests that they depended upon the invention of modes of

organization which required the subject or the person being tested to react only by very simple means. The time required to write out a long response, and the difficulty of scoring such a response, is a serious obstacle to its use in a group test. We may now describe briefly some of the most important devices which have been used for this purpose. Since the devices used in language tests differ somewhat from those used in non-language tests, we may consider them separately.

1. One of the earliest types of language test to be adapted to group testing is the completion test. The individual is required to supply a missing word, or part of a word, or group of words, in a printed passage, to supply part of the letters of a word or to supply a number in a series. This type of test has been used not only in intelligence scales, but also with success in educational tests. In both cases, but particularly in educational tests, the interpretation of the individual's response is more or less in doubt. If a person fails to supply correctly the missing item, we cannot be sure whether this is due to his lack of the information, which is required to supply it, or to his inability correctly to solve the puzzle which is represented in such a problem. This ambiguity is not so serious a difficulty in intelligence tests as in educational tests, since in either case intelligence may be the capacity which is required.

2. A second type of tests which is very common is the alternative or multiple choice type. This is represented, for example, in the *yes-no*, the *right-wrong*, or the *same-opposite* tests. In all of these cases the individual is required simply to make a choice between alternatives. In other cases a series of possibilities is represented numbering from three to five, and the individual is required to make choice among them. The multiple choice test may be illustrated by the analogies test. Examples of this may be found in

Army Alpha, which was reproduced on page 116. Another illustration is the classification test, or a test such as the following, which appears in the National Intelligence Test: "Underline the words in parenthesis which tell what the thing designated by the first word always has. *Lake* (*fish, salt, sand, shore, water*)." This test requires a good deal more intellectual activity than do some of the multiple choice tests. Another type requires the individual to check one of a series of answers which is the correct answer to a preliminary statement. Test III of Army Alpha is an illustration of this type.

Two types of questions arise concerning the alternative or multiple choice types of test. The first is concerned with the more general psychological significance of the reaction which is required by it. The second is a technical question regarding scoring, which will be discussed in the next chapter. The first question is this: How does the grasp of a fact which is sufficient to enable one to judge whether a statement concerning it is right or wrong compare with that grasp which is necessary to enable one himself to make a statement concerning it? Take a random illustration. The ability to give a correct answer to the question, "Does manual labor always result in cerebral hemorrhages?" gives very little information about the individual's knowledge of relation between physical exertion and disturbances of the circulation. One may reply that the purpose of the intelligence test is not to determine what the individual's information or ability to grasp a particular fact is, but rather to establish the relationship between his capacity and that of another person. For such purposes as general comparison, tests organized in this way have proved themselves to be very useful. It is worth calling attention to the fact, however, that they are not analytical. They do not enable the examiner to determine precisely and in detail the intellectual equipment of the individual.

3. A third device requires the individual to designate a superfluous part. This method has been used most prominently in the Pressey Mental Survey Scales, Cross-Out Tests. For example, a disarranged sentence is presented with one extra word. The subject is required to rearrange the sentence in mind and in this way determine what the excess word is and then cross it out. In another test a group of words is presented, all of which designate objects of the same class, except one. This one is to be crossed out. This type of test, if carefully planned, may necessitate a careful examination of the subject-matter of the test and real thought on the part of the examinee.

4. A fourth type requires that a series of items be arranged in rank order. This has been applied most often to moral judgment tests, or ethical tests, in which the individual is required to arrange in order of their seriousness a number of misdemeanors.

5. A fifth type may be classed with the fourth in that it raises somewhat similar questions. It requires that two series of words be matched in pairs. An example of this type of test is the "matching proverbs" test in the Otis Advanced Examination. It is hard to tell how far one's success in passing tests of this sort is due to his ingenuity in experimenting with the arrangement of the items, and how far it is due to his ability to comprehend the relationships which are involved. Again we may say that so far as securing a general measure of intelligence is concerned this question may not be important; but the distinction which has just been raised is important in its bearing upon the effect of training or practice in the ability to pass a group test. This practice effect has sometimes been found to be considerable, running about ten per cent or more. It seems likely that the practice effect consists largely of an increase in the ability of the individual to handle the mechanics of the test. Practice,

for example, may enable one to hit on certain devices, such as eliminating the obviously improbable answers and then examining the rest. If this is the case, differences in practice with this kind of material must be taken into account whenever we compare groups of individuals, or individuals who are likely to have had different opportunities to secure this practice.

6. A sixth type of test requires that the parts be rearranged so as to make sense. This method is used in the test which requires the words of a sentence, which have been placed in random order, to be rearranged so as to make sense. It appears in the Binet scale, and has been used in a good many group tests. It is at once a test of ingenuity and of the familiarity with the subject-matter of the sentence.

4. Modes of organization of the items of the non-language test

At least four of the types of examples which are used in the non-language test are analogous to the types which are used in language group tests. For example, the completion form is very commonly employed. The completion test is used with pictures in which one part is omitted, which the child is to draw in. It is also used with a series. An example of this is the so-called "X.-O" series, in the Army Beta test.

The alternative or multiple-choice test is also used. The opposites test is represented graphically by means of drawings of objects. The classification test is employed by showing drawings of objects in place of words.

Rearranging the parts of a series so that they shall be in an order which shall have sense is represented by the "Foxy Grandpa" series of pictures in the Performance Scale which was used in the army. Other series of pictures of a similar sort have been used.

The test which requires a superfluous part to be crossed out is represented by the commonly used absurdity test.

This consists of a series of pictures, each of which contains some part which does not belong in the picture.

In addition to these types of tests, which are analogous to those used in the language scales, are some that are peculiar to the non-language test. For example, the directions test is a very common one. This is illustrated by the first test in Army Alpha, which is, in fact, a non-language test. Many of the tests in the Dearborn Scale are directions tests, and the same is true of the Cole-Vincent Test for School Entrants, and a number of other primary tests.

A common form of test is the one which requires the recognition or combination of figures of different shapes. In some cases the figures are to be combined in order to make another figure. In some cases the form of a figure is to be reproduced in a drawing, and in other cases figures are to be matched. The cube analysis test of Army Beta is somewhat related to these tests.

Finally, we may mention the tests which require the interpretation of pictures. An example of this type is a picture of a boat which the child is required to interpret so as to tell whether the boat is moving or still.

The same critical questions which were raised in regard to the types of tests used in the language scales may be applied to these types when they are used in the non-language material. We have recently passed through a period which has been very fruitful in the invention of new devices for group testing. It is now desirable that we should have a much larger amount of critical experimental evaluation of these methods than has characterized the period of rapid development.

Chapter XI

TECHNIQUE AND THEORY OF MENTAL TESTS

III. Problems Relating to Scores and Norms

1. Mental test scores: the raw score

THE score in a mental test is, of course, a numerical quantity. The meaning of this quantity, however, depends upon the nature of the material of which the test is composed, and the way in which the material is organized. The raw score is the expression of the achievement of the individual in terms of the unit of which the scale is composed. The raw score has no significance in itself. The same raw score may mean a different thing in the case of different tests, according to the unit which is employed, or to the conditions of the test. One exception to this statement is to be found in the case of mental age, provided we include this among the raw scores. The raw score takes on significance as it is translated into comparative or relative measures. The way in which this is done will be considered after we have mentioned a number of illustrations of raw scores.

The raw score may consist, in the first place, in a numerical statement of the amount which is accomplished within a given time limit. An illustration of such a raw score is that which is obtained by counting the number of letters which a person crosses out in a printed text. Another score of similar character expresses the number of substitutions which a person makes, as in the case of the digit-symbol test. It is obvious that these raw scores are affected by the character of the material which is used.

The amount done when a time limit is not imposed is a type of raw score which, as we have already seen, may be

regarded as a measure of power. Illustrations of this sort of score are obtained from tests of span of attention. For example, the test used in the Binet scale which requires the individual to reproduce a list of numbers which is spoken to him measures the limits of capacity in this kind of performance at the time of the test. It is not affected primarily by the speed with which the individual is required to respond. A test in which the items are arranged in increasing difficulty to a point beyond the capacity of the individuals taking it, when the time is not limited, may be called a power test.

The score of a speed test may be expressed in terms of the amount done in a given period of time, or the time required to do a given amount. The amount performed in a given time can be used in a group test, but the time required to do a given amount can be used most conveniently in an individual test. This is because it is not easy to record the time which is occupied by various individuals of a group in performing a given set task.

The time score has an advantage over the amount score in that all the individuals who are given the test perform the same amount of work. If there are irregularities in the difficulty of the different parts of the test, these will affect equally the scores of all the individuals. If the time limit is used, however, some individuals may meet certain difficulties which others escape. The difference is not one of great importance if the items of the test are well graded.

Another type of raw score is given in terms of units discriminated. This type of score is illustrated in the sensory discrimination tests. For example, in the test for discrimination of pitch the unit of measure is the vibration frequency. The score of the individual is the least difference between the vibration frequencies of the two tones which can be discriminated, assuming a certain basic pitch as a

standard of comparison. In the case of weight discrimination, the unit difference might be expressed in fractions of an ounce, or grams, or any other unit of weight. In discrimination of the intensity of sound, the unit needs to be expressed in terms of the instrument which produces the sound.

The most common type of raw score in current use is the point score. The most common method of finding the point score is to add up the items of a test which the individual passes. In some cases a deduction is made for errors, and in other cases different parts of a test are given different weights. These procedures will be discussed later in the chapter. In some cases the point score is made up of such constituents as the number of moves which are made in passing a test, and the time which is taken. In all cases, the point score is quite obviously a raw score, in the sense that it is not self-interpretative. Its significance needs to be found by comparison with a standard.

The mental age may be regarded as another form of raw score, but it is different from the ones which have been mentioned in that it carries with it its own significance. This significance, however, is incomplete, unless a relative score, such as the I.Q., is found.

2. The accuracy of the score and the sources of error

We have already seen in our reviews of Spearman's critique of mental testing, in Chapter III, that he called particular attention to the problem of the accuracy of the scores in mental tests, and that he proposed methods for determining the accuracy of scores and for making allowance for errors. We may now proceed to an analysis of the sources of errors as they have been brought out in subsequent investigation.

A clear account of errors in test scores has been given by

Holzinger.[1] We shall follow his classification. He distinguishes five types.

1. Scale errors. These errors are inherent in the tests themselves. They consist in the selection of unsuitable material for the test, or in the imperfect gradation or arrangement of the material. The correction of scale errors can be accomplished by the improvement of the test itself.

2. Scoring errors. Scoring errors occur chiefly in "product scales," in which the pupil's product is compared with those which compose a scale. They are errors in judgment which occur in grading specimens. These errors can largely be eliminated in the usual type of mental test in which the pupil makes a response that can be scored objectively and in quantitative terms.

3. Response error. This is a very troublesome error. It is caused by the actual fluctuation in the pupil's response from one occasion to another, caused by changes in emotional condition, interest or effort. It is this error chiefly which is estimated by the reliability coefficient. It is avoided by using the composite result of a sufficient number of tests.

4. Sampling error. This error appears when we take the scores of a given group as representative of another group or of persons in general. It applies especially to the use of norms.

5. Sporadic error. "Sporadic errors are those due to arithmetical blunders in scoring, misunderstanding of test directions, time lost by the pupil with a broken pencil, etc. Such errors may be eliminated" (p. 281).

The most important of these errors, except for the use of norms, are those of scoring and those of response. If the scoring errors are eliminated, as they may be, our concern is with response errors. The interpretation of the score is determined by the amount of this error. Holzinger gives formulæ by which it may be estimated. It is important for the user of mental tests who does not have at his command

[1] Karl J. Holzinger, "An Analysis of the Errors in Mental Measurement," *Journal of Educational Psychology*, XIV (May, 1923), 278–88; see also the following for a more extensive and detailed list of sources of error: Percival M. Symonds, "Factors Influencing Test Reliability," *Journal of Educational Psychology*, XIX (February, 1928), 73–87.

the technique of estimating response errors to take two precautions. First, he should be particularly wary of conclusions drawn from repeated test scores which are designed to show the individual's progress. Second, in estimating a pupil's ability he should avail himself of the result of several tests, rather than of one alone.

The term "error" is here used in a broad sense to cover all sources of variation in the score. Strictly speaking, not all variations are due to error. If the score faithfully represents the actual performance of the individual at the time he takes the test it is a correct measure of that performance. There are at least three fundamental ways in which performance may be different on successive occasions. First, the individual may react differently to a situation which is externally exactly the same. This is true even of reflex acts, such as turning the eye toward a flash of light. Second, the external situation may be the same in general character but may be different in its particulars, producing a difference in response. Thus the elements of a test may vary though they are all of the same kind. We may call this a difference in sampling of acts which represent the same general kind of ability. Third, the individual may actually have changed from one time to another. The change may be due to growth and development, to education and training, or to deterioration and decline. Such changes commonly take a long time. Therefore, comparisons after long intervals are usually of a different order from those after short intervals.

The aim of mental tests is to secure as stable and constant measures of ability as possible. The measure of such stability and constancy is the reliability of the test. The usual method of measuring reliability is to repeat the test and correlate the scores on the two administrations. If the correlation is .90 or above the score on one administration of the test serves fairly accurately to predict what the

individual will do on another occasion. The sum total of the errors and of the variations in performance will not be large. This repetition should be made after a short interval to exclude actual permanent changes in the individual. These are not the fault of the test and should not be considered sources of unreliability.

The actual procedures in calculating reliability vary somewhat. In some cases the odd and even items of a test are correlated and the coefficient is corrected by the Spearman-Brown formula, because this procedure in effect shortens the test by one-half. In other cases the entire test is repeated, and in still others two forms of the test are correlated. In the first and third methods there is a different sampling of performances, whereas in the second exactly the same performance is required. This fact introduces somewhat different elements of variation. The coefficients, however, are roughly comparable and serve to indicate whether the test is, in general, satisfactory from the point of view of the accuracy or stability of its scores.

3. Treatment of wrong answers

Except in special cases the wrong answers in a mental test are disregarded. The score is the total number of correct answers. There are certain cases, however, in which it seems theoretically desirable to take some account of the errors as well as of the correct responses. These are the cases in which the individual is required to designate only which of several answers is the correct one, the so-called *multiple-choice*, or *yes-no*, or *right-wrong* tests. The necessity of taking account of the errors comes from the fact that it is possible for the individual to obtain the correct answer in a certain number of cases by pure guessing. If the score is to represent exactly the individual's knowledge or capacity, such right answers ought to be discounted. The practice has

been to determine how much the score should be discounted by calculating from the number of errors the number of right answers which the individual probably got by guessing.

The theory and the resultant practice may be set forth by an illustration from the *yes-no* type of test. The theory may be illustrated thus: Suppose that we were in possession of all the facts concerning an individual's response to a test. Suppose, then, the following situation exists. There are twenty items in a test. The individual knows ten of the items. He therefore passes these ten correctly because he knows the answers. He attempts six more items, but guesses on all of them. According to the theory of chances he would get three right and three wrong. (This, of course, would only be true in the long run, and not in each particular case, but only in a certain proportion of them.) The number of right answers would then be thirteen, the number of wrong answers three. If, now, we started only with a knowledge of the number of right answers and of the number of wrong answers, we could work back to the true score by subtracting the wrong answers from the right answers. This would give us the number which the individual got right because he knew the answers. This number would, of course, be ten. We may express this procedure for finding the correct answers in the following formula: True Score = Right − Wrong.

This procedure has been adopted almost universally in tests of the alternative type. It has been vigorously criticized from various points of view. The first criticism is based upon the theory of chances. The procedure assumes that if an individual guesses on a certain number of the items of the test, he will guess right as many times as he will guess wrong. As was parenthetically remarked in the previous paragraph, this would be true only in the long run. It would not be true in a large proportion of the individual cases. This is because the number of items on which the

person guesses is so small. If there were one hundred, for example, the proportion between right and wrong guesses would be close to 50 per cent in nearly all cases. This objection, then, holds chiefly when the number of items in a test is small and is not serious when the number is large.

The second objection is a psychological one. We cannot assume that all the wrong answers are guesses, nor that a person guesses on as many answers which he got right as on those which he got wrong. West shows, in an experimental study of the alternative response type of test, that the subjects may not guess right the same number of times that they guess wrong, and that the score which is obtained from the formula is not the same as the true score derived when we know the items on which the individuals actually guess.[1] West had a class of college students take an opposite test consisting of fifty items. He had the individuals tell him all of the items on which they were confident of the answers and all on which they guessed. He found that the scores obtained by taking the testimony of the subjects as to which items they knew and those obtained by using the formula

$$S\ (\text{Score}) = R\ (\text{Rights}) - W\ (\text{Wrongs})$$

did not agree closely. This was partly because they got more right than wrong when they guessed or thought they guessed. Apparently they did not always know when they were guessing.

As a substitute for this theoretical method of determining what deduction should be made for errors, Thurstone proposes that the proper deduction should be reached empirically. He suggests a formula by which may be determined what deduction for errors will give the highest correlation between the tests and the criteria.[2] The purpose of the

[1] Paul V. West, "A Critical Study of the Right Minus Wrong Method," *Journal of Educational Research*, VIII (June, 1923), 1–9.

[2] L. L. Thurstone, "A Scoring Method for Mental Tests," *Psychological Bulletin*, XVI (1919), 235–40.

procedure is not primarily to allow for guessing, but to find the proper relative weight to give to speed and accuracy. Thurstone gives data to show that a deduction for errors which is determined by this empirical method will give a good correlation with the criterion, whereas another type of deduction for errors will give a much lower correlation. This procedure is probably preferable to the use of the R − W formula, but it is cumbersome to use and probably would not be satisfactory in all cases.

Holzinger [1] has shown that all necessity for allowing for guessing disappears when all the individuals are given opportunity to attempt all the items of the test, since under these circumstances there is a perfect correlation between the scores consisting of the right answers and of the rights minus the wrongs. This procedure, of course, cannot be followed in a test in which a time limit is imposed.

In spite of the objections to the calculation of the score on alternative or multiple-choice tests by deduction for errors, the results of different methods of scoring favor such deduction for alternative tests, because of the simple formula based on the theory of chance. A comprehensive summary of studies of the problem is given by Lee and Symonds.[2] The formula which includes multiple-choice tests is

$$S = R - \frac{W}{n-1},$$

n being the number of choices offered. It is doubtful whether it is worth while to make a deduction except for the true-false or yes-no type of test. It seems best, also, to instruct the testees not to guess. It seems desirable, wherever

[1] Karl J. Holzinger, "On Scoring Multiple Response Tests," *Journal of Educational Psychology*, XV (October, 1924), 445–47.

[2] J. Murray Lee and Percival M. Symonds, "New-Type or Objective Tests: A Summary of Recent Investigations," *Journal of Educational Psychology*, XXIV (January, 1933), 21–38.

convenient, to use another type of test in place of the alternative type. The increase of the number of choices from two to three, four, or five decreases the chance of error from guessing proportionately. The definiteness and the ease of scoring of the multiple choice test makes it a very useful one. It is particularly useful when our purpose is merely to discover the relative grasp of the subject or the relative intelligence of individuals of a group. If we wish to determine absolutely the amount of information possessed, however, the test has a definite limitation. This limitation grows out of the fact that it requires a much more thorough grasp of a subject to give an item of information independently than is required to designate the correct answer out of a number of possible answers. If a more independent grasp of information is to be measured, it is better to use the completion type of test, or some other type which requires the individual to supply the answer himself instead of indicating a choice among answers.

4. Weighting test scores

Scores are weighted in order to modify the share which the raw scores of the individual items of a test have in the total score, or to modify the share which the raw score of one test has in the composite score of a group of tests. In some cases weighting is introduced to equalize the share of various items or various tests, and in other cases to make the share of the items or of the tests unequal.

Weighting for the purpose of making the share of various tests of a scale equal is commonly represented in the following situation. Suppose that a scale consists of five tests. Suppose further that these tests contain an unequal number of items, for example: test 1, 30; test 2, 20; test 3, 10; test 4, 15; and test 5, 25. If the difficulty of the various tests is

scaled alike, it is clear that test 1 will contribute three times as much to the composite score as test 3, and twice as much as test 4. Unless we regard test 1 as more important than the other tests in this proportion, the use of the raw scores will throw the scale out of balance. This is usually prevented by multiplying the score on each of the tests by a factor which will make the total possible score on each test about equal to that on the others. Since the difference between the fifth and the first test is small, we might disregard it and correct only the scores on tests 2, 3, and 4 by making the total possible score on each thirty; thus the score of test 2 would be multiplied by $1\frac{1}{2}$, the score of test 3 by 3, and the score of test 4 by 2.

It is not altogether clear that the weighting of the tests of a scale so as to equalize their share in the total score is necessary or desirable. The correlation between the total scores of scale A, in which the individual tests are weighted in this way, with the total raw scores, is given in the Army Report (page 340). The correlation with one group of nine hundred men between the weighted and unweighted scores was .994. With another group of three hundred men it was .93. Weighting of this sort appears to make little difference in rank. This is what we should expect theoretically. Equalizing the share of the different tests is important only if the tests measure relatively different and distinct mental capacities. If they measure the same capacities the existence of an unequal share of the different tests in the total scores is not serious. While it is, of course, true that the content or subject-matter of the various tests of our intelligence scales is different, it is questionable whether they measure fundamentally different mental processes. We have no good evidence that they do. It appears, at any rate, that the distinction between what is measured by the various particular tests is not sufficiently clear to

warrant the refinement in method which is represented by weighting their scores.

The second aim of weighting is to make the scores of the items, or of tests, unequal. This type of weighting is based upon the assumption that the importance of the items or of the tests is not the same. The variation in importance is in general based upon one of two facts. The first is the difference in difficulty of the various items of the test. It is sometimes assumed that the more difficult item should be given greater weight than the easier items. The items of the test are therefore multiplied by a factor which is proportional to their difficulty. The second basis for estimating the importance of items, or of tests, is their correlation with the criterion. When this basis is used, a test is weighted by multiplying by a factor which is roughly proportional to the correlation between that test or item with the criterion.

Weighting for the purpose of making the share of tests in the total score unequal is less commonly used than it was a number of years ago. The procedure of determining the weights and of applying them to the scores is cumbersome, and empirical studies do not seem to indicate that the resulting score is better than the raw score.

A series of high correlations between weighted and unweighted scores in a standardized educational test in algebra were found by Douglass and Spencer.[1] Four correlation coefficients ranged from .98 to .996. Holzinger[2] also reports high correlations between weighted and unweighted scores and points out that when this correlation is much higher than the self-correlation (reliability coefficient) of the test the use of weighted scores is unjustified.

[1] Harl R. Douglass and Peter L. Spencer, "Is It Necessary To Weight Exercises in Standard Tests?" *Journal of Educational Psychology*, XIV (February, 1923), 109–12.

[2] Karl J. Holzinger, "An Analysis of the Errors in Mental Measurement," *Journal of Educational Psychology*, XIV (May, 1923), 278–88.

These results would seem to agree with a common sense analysis of the problem. Assuming that the tests of a scale are arranged in ascending order of difficulty, and each item is assigned a score of 1, the person of high ability will always make a score superior to that of a person of low ability. The assigning of greater weight to the difficult problems as compared with the easy ones will make a difference in the relative size of the scores, but not in the order of the scores. Since the order or ranking of the scores is the important thing, and we seldom or never attempt to calculate the ratio between two scores, the unweighted score is as serviceable as the weighted score.

5. Measures of relative standing

In the previous section we have discussed the raw score and the method of obtaining it. It has been said that the raw score in itself has no meaning. It needs interpretation. The only original score, or raw score, which carries its own interpretation is the mental age. Even this, however, has a limited interpretation. We may now consider, first, the further methods of interpreting the mental-age score, and second, the methods of interpreting point scores. In general, these methods consist of turning the raw scores, which are in absolute terms, into relative scores.

The first relative score which was used to interpret the mental age was a difference, namely, the difference between the mental age and the chronological age. Thus, if an individual has a mental age of twelve and chronological age of ten, his intellectual superiority would be represented by a score of twelve minus ten, or two years. This means that an individual's intellectual development is two years beyond what we should expect from his chronological age.

As was remarked in discussing the Binet scale and its development, it was soon discovered that the significance of

this mental age-chronological age difference was not the same for the different stages of the child's life. A year's difference was found to be more significant, when the measurement was made by the Binet scale, in the case of the young child than in the case of the older child. To avoid variation in the meaning of the measure, another measure was used which consists of a ratio rather than a difference. This is the familiar intelligence quotient, which is the ratio between the mental age and the chronological age. Thus a child whose mental age is twelve and chronological age ten would have an intelligence quotient of twelve divided by ten, or 1.20 (usually written 120). The same quotient would represent the intelligence of the child whose mental age is six and chronological age five. The intelligence quotient of these two children would be the same, while the difference between the mental and chronological age would be twice as much in one case as in the other.

Since this ratio, which has been found empirically to work well with the Binet scale and its modifications, has been used in some instances with other types of scales, it is desirable to inquire into the assumptions upon which it is based and into the conditions under which it may legitimately be applied.

The fundamental requirement of a relative measure of intelligence, as already suggested, is that it shall remain constant throughout the period of mental development. By constancy is here meant that a particular I.Q. shall have the same significance in all the ages to which it is applied. It does not refer to the constancy of the I.Q. of any particular individual. That is another aspect which must be discussed separately. The use of the I.Q. as a measure, however, means that a particular I.Q. shall have the same significance at six years that it has at ten or twelve years. To put it another way, if we should make a distribution of an unselected group of individuals at a series of ages, and should

calculate their I.Q.s, the individuals at corresponding points in the various distributions should have the same I.Q. For example, individuals at the lower quartile of each distribution should have the same I.Q. as the individuals at the lower quartiles of all the other distributions. To put it another way, the variability or range of the I.Q.s at the various ages should be the same. This requirement seems to be met substantially by the I.Q.s calculated from the Binet scale. What is the explanation for this constancy of the I.Q.?

We may analyze the situation most readily by the graphic method. As the writer has pointed out in another place, there are two statistical facts which are involved in the problem. These are first, the form of the age progress curve, and second, the relative distribution of the scores in the succeeding ages. If either of these two factors is constant, the other one may vary in such a way as to make the I.Q. valid and comparable from year to year. That is, if the age progress curve is a straight line or if the increments from year to year are the same, the distribution of scores in succeeding ages may be such as to render the I.Q. valid. On the other hand, if the distribution of scores from year to year is the same, the form of the age progress curve may be such as to render the I.Q. valid.[1]

Fig. 14 illustrates the case in which the yearly increments are uniform but the spread of the distribution in the succeeding years increases uniformly and proportionately. It also involves the assumption that the growth curves have their origin at birth. The upper line represents the curve of the individuals of median ability; the lower line represents the curve of individuals at some lower level, in this particular

[1] Frank N. Freeman, "The Interpretation and Application of the Intelligence Quotient," *Journal of Educational Psychology*, XII (January, 1921), 3–13.

case with an I.Q. of .66. The mental age of the individuals at any point in the lower curve is found, of course, by finding the point on the median curve which is on the same horizontal level with it and then by projecting downward to find the age which corresponds with this point. Thus, the in-

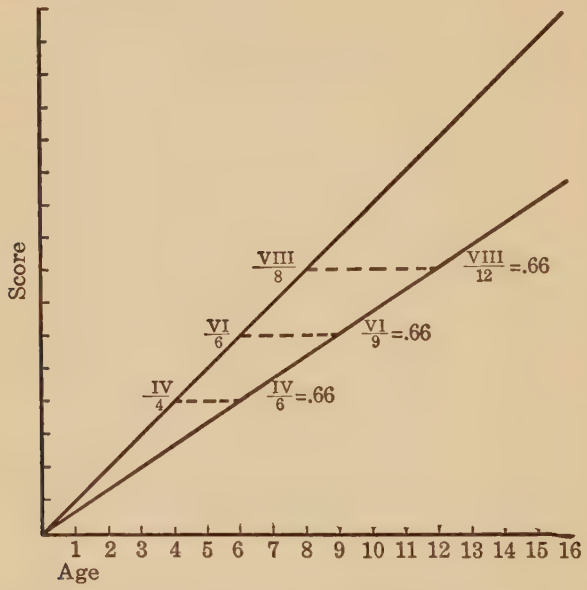


FIG. 14. HYPOTHETICAL GROWTH CURVES TO GIVE A CONSTANT I.Q.

dividual on the lower line of development at age nine has a mental development which corresponds to that of the median individual of age six. The intelligence quotient of this individual is, then, six divided by nine, or .66. In the same way the I.Q. of any individual at any point on the lower line may be found, and it will be seen that it is always the same. That this is so can be demonstrated geometrically on the principle that the sides of similar triangles are proportional.

The second case is illustrated in Fig. 15. The upper curve, as before, represents the mental growth of the median individual and the lower curve the growth of an individual of inferior capacity. In this case, the feature which is constant from year to year is the distribution of the scores. This is represented by the vertical distance between the two lines. It will be seen that this is the same in succeeding parts of

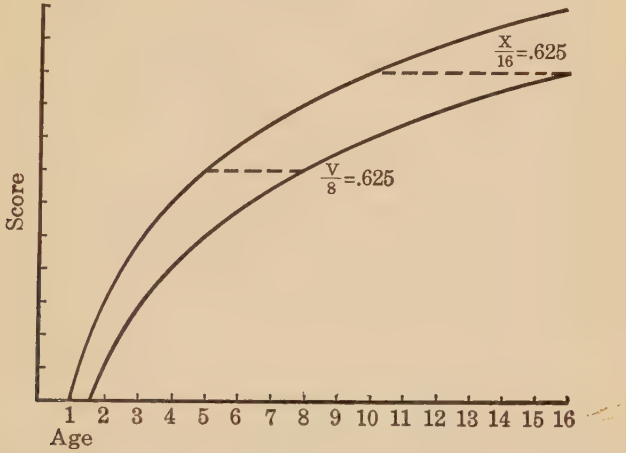


Fig. 15. Hypothetical Growth Curves to Give a Constant I.Q.

the curves. The curves of development, however, are not drawn as straight lines, but as logarithmic curves. That is, the heights of the curve at the points above the base line representing the various ages are the logarithms of these ages. For example, the logarithm of two is .301, of three, .477, of four, .602, of five, .698, and of six, .778. This produces a curve which rises sharply in the early ages and more and more slowly in the later ages. If, now, we calculate the intelligence quotients of various individuals which are represented by the lower curve, in the same fashion as for the pre-

ceding figure, we find that these quotients are the same at the various ages. In other words, an age development curve which has the logarithmic form, assuming that the distribution at succeeding ages is the same, will give a constant I.Q.

Another way of expressing the condition underlying the constancy of the I.Q. is to say that the overlapping of the scores of succeeding ages must be a proportionately increasing one from year to year. This increase in the overlapping of scores may be due to the diverging lines of yearly development, on the one hand, or to the decreasing rate of mental growth, on the other hand. An examination of the two figures will show that this increase in overlapping is regular and increases proportionately.

It was the observation of this increase in the overlapping of scores from year to year which first called attention to the necessity of a ratio to express intelligence in the case of the Binet scale. This overlapping was not analyzed, however, in order to determine whether it was due to a decreasing rate in mental growth, to increasing range of distribution, or to a combination of the two. The issue has perhaps most clearly been set forth by Woodrow.[1] Woodrow has drawn his curves so as to represent both a decreasing rate of maturity and an increasing distribution from year to year. These curves have obviously been drawn empirically, however, in such a way that the combination of these two factors will produce a constant intelligence quotient. The data of the Binet scale give us a measure of the amount of overlapping of scores from year to year, but since they are represented in terms of mental age, or because they are standardized by age, they give us no means of determining which of these two factors or what combination of them is at the basis of the constancy of the I.Q.

[1] Herbert Woodrow, *Brightness and Dullness in Children*, pp. 46–48. Philadelphia: J. B. Lippincott Co., 1923.

The use of the I.Q. assumes that the conditions of rate of mental growth at successive ages and relative variability at successive ages, in combination with each other, will yield I.Q.s which will have the same standard deviation at successive ages. For the American revisions of the Binet scale this is approximately true. Why it is true, however, is not analyzed because the scores are in terms of mental age which does not permit us to derive a curve of mental growth. We merely have the resultant fact that the S.D.s are fairly constant.

The approximate equality of the S.D.s, however, apparently does not tell the whole story. There is a variation at the extremes of ability such that the Stanford-Binet I.Q.s of gifted children tend to rise whereas those of feeble-minded children tend to fall.[1] This variation indicates that the I.Q. gives a more uniform measure of ability from age to age in the middle part of the scale than toward the extremes.

An attempt to develop a mode of calculation of relative scores which give more constant scores from age to age at the extremes has been made by Heinis.[2] Heinis first worked out a formula for a growth curve from the scores on the fifteen tests of the Vermeylen scale made by sixty normal children from six to twelve years of age. This curve of growth is expressed in terms of "mental growth units." After a child's mental age has been found it is translated into these units. His personal constant, P.C., is then calculated by finding the ratio between his score in mental growth units

[1] Psyche Cattell, "Constant Changes in the Stanford-Binet IQ," *Journal of Educational Psychology*, XXII (October, 1931), 544–50; and F. Kuhlmann, "The Results of Repeated Mental Re-Examinations of 639 Feeble-Minded over a Period of Ten Years," *Journal of Applied Psychology*, V (September, 1921), 195–224.

[2] H. Heinis, "A Personal Constant," *Journal of Educational Psychology*, XVII (March, 1926), 163–86. *See also* Katherine Preston Bradway and E. Louise Hoffeditz, "The Basis for the Personal Constant," *Journal of Educational Psychology*, XXVIII (October, 1937), 501–13.

and the average score for his age as also expressed in the mental growth units. A series of tables for finding the P.C. given the mental age and chronological age has been made up by Hilden.[1] This scheme assumes the correctness of the given growth curve and also assumes that the ratios of mental growth units are constant from age to age. If the P.C.s are found for other scales than the one on which the original calculations were made, it assumes that the mental growth curves of these scales are the same as that obtained by Heinis. The method involves an extra step. At best it serves as a correction of the I.Q. by a somewhat roundabout method. At worst it is more complicated without being better.

Empirical checks on the Heinis P.C. are not entirely conclusive. Kuhlmann and Hilden,[2] for example, find that the median change in P.C. is less than that in I.Q. Cattell,[3] on the other hand, finds the P.C. somewhat more constant for dull children but much less constant for bright children.

When we come to deal with point scales, we can determine independently the form of the age curve and the distribution of scores from year to year. It is not necessary to discuss the fact in detail at this point.

We may sum up the matter by saying that the age-growth curve seems to approach much more nearly a straight line than a logarithmic curve, within the limit of those ages for which a particular test is well suited, and up to the period of adolescence. So far as the distribution is concerned it seems to increase somewhat from year

[1] A. H. Hilden, *Table of Heinis Personal Constant Values.* Minneapolis: Educational Test Bureau, 1933.

[2] Arnold H. Hilden, "A Comparative Study of the Intelligence Quotient and Heinis' Personal Constant," *Journal of Applied Psychology*, XVII (1933), 355–75.

[3] Psyche Cattell, "The Heinis Personal Constant as a Substitute for the IQ," *Journal of Educational Psychology*, XXIV (March, 1933), 221–28.

to year, but not enough to make the intelligence quotient constant.

Rand has shown, in the article cited on page 298, that in fact, there is not a combination of decreasing mental growth and increasing range of distribution in the case of most point scales, so as to produce the proportional increase and overlapping from year to year which is necessary as a foundation for a valid I.Q. The I.Q. is not a suitable measure, then, for use with the ordinary point scale.

Examples of the variations among I.Q.s derived from different tests and a suggested method of equating the I.Q.s on different tests are given by Miller.[1] The method of equating consists of a chart by means of which the I.Q. on a given test may be translated into terms of the standard deviation of the I.Q. or the reverse. This method serves to equate scores on different tests, but not to equate scores for different ages of the same test.

Another ratio to express relative intellectual capacity, somewhat similar to the intelligence quotient, is the coefficient of intelligence. This ratio, which was first used by Yerkes, Bridges, and Hardwick in their point scale, is the ratio between the point score of the individual and the point score which is the norm for his age. Thus if the norm for a given age is 100, and the individual makes a score of 80, his coefficient of intelligence is .80.

The coefficient of intelligence has not been very widely used and its relation to the intelligence quotient has not been much discussed. A moment's thought, however, will show what that relationship is. If the coefficient of intelligence is to be a valid relative measure of intelligence, it must, like the intelligence quotient, have the same significance from age to

[1] W. S. Miller, "The Variation and Significance of Intelligence Quotients Obtained from Group Tests," *Journal of Educational Psychology*, XV (September, 1924), 359–66.

age. That is, it must be constant. The condition necessary to make the coefficient of intelligence constant is that the spread of the distribution in succeeding ages shall increase proportionately. The coefficient of intelligence, unlike the intelligence quotient, is not affected by the form of the age progress curve, except as this may affect the spread of the distribution. The only case in which both the intelligence quotient and the coefficient of intelligence can remain constant is the one in which the age progress curve is straight and the spread of the distribution increases regularly and proportionately.

The variability in the I.Q. and the C.I. in different tests or in different ages of the same test is amply illustrated in a paper by Gertrude Rand.[1] She shows that, in general, the variability of I.Q.s increases with age while the variability of C.I.s decreases with age. The latter phenomenon is, of course, due to the fact that the variation in scores with increasing age is not proportional to the increase in the scores. The increase in spread of I.Q.s is due to the fact that the negative acceleration of the curve of mental growth and the increase in the variability of test scores in succeeding ages combined are not sufficient to keep the I.Q. constant. Rand's data also reveal an enormous difference in the spread of I.Q.s in different tests.

The case seems rather paradoxical. First we find that the I.Q. is approximately constant in the case of the Stanford-Binet scale. From analysis we conclude that this constancy implies that there is a diminishing rate of growth or an increasing divergence in abilities or both. On the other hand, it appears that the I.Q. is not constant for many point scales, nor comparable among the various scales. Furthermore, to

[1] Gertrude Rand, "A Discussion of the Quotient Method of Specifying Test Results," *Journal of Educational Psychology*, XVI (December, 1925), 599–618.

anticipate, we shall find that point scales do not give us the conditions which we have found by analysis to be necessary to give a constant I.Q. The Stanford-Binet seems to indicate one type of development and the point scales another.

The solution of the paradox is to be found in the fact that the form of mental-growth curves depends not only on the fundamental nature of mental development itself, but also on the characteristics of the scale which is used to measure it. Thus, some scales will show a retardation in mental growth at a particular period while others show a uniform rate of advancement at the same period. We cannot draw universal conclusions from the results of a single scale regarding either the applicability of a particular type of score, such as the I.Q., or the form of the curve of mental growth.

Another relative score which looks superficially like the I.Q. or the coefficient of intelligence, but which is based on a fundamentally different assumption, is the index of brightness, or the I.B., first used by Otis. The index of brightness is found by calculating the difference between the individual's score and the norm for his age, and then, according as this difference is plus or minus, adding it to, or subtracting it from 100. Thus if the norm is 90, and the individual score is 97, his index of brightness is $7 + 100$, or 107. If the individual score is 83 the I.B. is 93. It is obvious that a given amount of superiority of a score in points, or inferiority in points, is given the same significance at various ages by this method of calculation. Ten points above the median at age six means exactly the same thing as ten points above the median at age twelve. This is fundamentally different from the principle underlying the coefficient of intelligence. It presupposes, if the I.B. is to be constant, that the spread of the distribution in succeeding ages is identical in terms of points, and that the curves of age progress of individuals of various degrees of capacity are parallel.

The validity of this measure is not affected by the form of the age-progress curve. It is therefore possible that it may be consistent with the I.Q. It is improbable that it should be so consistent, however, since this would be true only in case the constancy of the I.Q. is based wholly upon the form of the progress curve and not to any extent upon the relative spread in distribution from age to age.

The relative scores which have thus far been described raise problems concerning the facts of mental growth, and their validity depends on the correctness of certain presuppositions concerning such growth. The assumptions of the I.Q. and the P.C. are especially complex because, since they employ the mental age, they involve the determination of the ability of children of given ages in terms of the attainment of children at other ages. The C.I. and the I.B. involve opposite assumptions to each other concerning the distribution of mental abilities.

These difficulties are avoided by relative scores which express the ability of an individual in terms of the distribution of scores of his own age group. One method of doing this is to express scores in terms of percentile rank. The percentile rank represents the position of the individual in a group of one hundred. A percentile rank of 25 means that there are twenty-five individuals out of one hundred whose scores are lower than that of the individual in question. This kind of score is comparable in reference to different tests with different kinds of raw scores.

The percentile rank has the advantage of simplicity and convenience. It has the theoretical defect, however, that it assumes the rectangular distribution of abilities instead of the normal distribution, to which the distribution of abilities in fact more nearly conforms. To illustrate, according to the normal distribution, the lowest 10 per cent of a group of individuals would cover a much wider range of the scale,

which is represented by the base line of the distribution, than would a 10 per cent group near the center of the distribution. By the rectangular distribution, however, the 10 per cent at the low end or the high end of the scale would cover the same distance as the 10 per cent in the middle. As a consequence, the percentile method is not suitable for precise scoring.

The kind of score which expresses the individual's attainment in relation to that of other individuals most accurately is the standard score. This score takes the standard deviation (often called sigma) of the distribution of the scores of the group to which the individual belongs as the unit of comparison. In the case of mental tests the group consists of the individuals of the same age as the person being tested.

The first step in finding the standard score is to calculate the S.D. (standard deviation) of the scores of the age group. This is the square root of the mean of the squares of all the deviations of scores from the mean of the group.

$$S.D. = \sqrt{\frac{\Sigma d^2}{n}}$$

The next step is to find the difference between the individual's score and the mean score of the group, expressed as plus or minus. The final step is to divide this difference by the S.D.

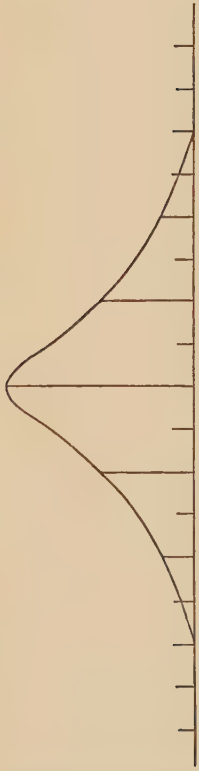
$$\text{Standard Score} = \frac{M-S}{S.D.}$$

Standard scores are awkward to handle because they include minus quantities and fractions. To avoid this the T-score was devised. It is found by expressing the mean score as 50 instead of 0 and dividing the S.D. by 10, which is, in effect, multiplying the deviation by 10. Thus, a mean

score will be 50, and a score which is 1 S.D. above the mean will be 60 (50 + 10), and one which is 1 S.D. below the mean will be 40 (50 − 10). The relation of the various scores to each other may be seen in the accompanying table of equivalent scores.

EQUIVALENT MEASURES OF BRIGHTNESS [1]

(Assumption of Normal Distribution)



Standard Scores	Percentile Ranks	New Stanford Rev'n I.Q.s	T-scores (12-yr. olds)
+4 sigma	99.99+ P.R.	164 I.Q.	90 T-score
+3.5	99.98	156	85
+3	99.86	148	80
+2.5	99	140	75
+2	98	132	70
+1.5	93	124	65
+1	84	116	60
+0.5	69	108	55
Median	50	100	50
−0.5	31	92	45
−1	16	84	40
−1.5	7	76	35
−2	2	68	30
−2.5	1	60	25
−3	0.14	52	20
−3.5	0.02	44	15
−4	0.01−	36	10

[1] This table was prepared by Dr. F. A. Kingsbury whose permission to use it is gratefully acknowledged by the author.

6. Measures of the relation between intelligence and achievement

Intelligence tests are assumed by theory to measure native capacity. Educational tests are assumed to measure the actual achievement which the individual makes. This is the product of his native capacity and the training which he has received, plus certain character traits and general environmental influences. It was inevitable that the attempt should be made to bring these two measures into relationship to one another in order to determine the degree to which the individual's achievement corresponds to his capacity. This was first done, so far as the writer is aware, by Buckingham and Monroe, in connection with their Illinois Examination.[1] These authors call the measure of relative achievement the "achievement quotient," or A.Q., and find it by dividing the achievement age by the mental age. Perhaps the most elaborate use of such a quotient as this has been made by Franzen.[2] Franzen follows substantially the same procedure as was followed in the Illinois Examination. He first finds the subject ratios of the various individual school subjects. These are the ratios between the subject ages and the mental ages. The average of these subject ratios he calls *accomplishment ratio* (Acc. R.). The accomplishment ratio, then, is the same as Buckingham and Monroe's achievement quotient. Franzen has since abandoned the practices here referred to.

The interpretation of the achievement quotient or the

[1] Walter S. Monroe, *The Illinois Examination.* University of Illinois Bulletin, Vol. XIX, No. 9. Urbana: University of Illinois, 1921; W. S. Monroe and B. R. Buckingham, *The Illinois Examination I and II, Teacher's Handbook.* Bloomington, Illinois: Public School Publishing Co., 1920.

[2] R. H. Franzen, "The Conservation of Talent," in Lewis M. Terman and Others, *Intelligence Tests and School Reorganization,* Chapter IV. Yonkers-on-Hudson, New York: World Book Co., 1922.

accomplishment ratio has not always been clear, and its use has sometimes led to interpretations which were absurd. For example, the accomplishment ratio is sometimes described as measuring the relation between one's achievement and one's capacity. According to this definition it would be impossible to have an accomplishment ratio above 100, since one could not achieve beyond his capacity. We do as a matter of fact, however, find a large number of accomplishment ratios above 100. We escape this particular difficulty if we describe the accomplishment ratio as a measure of the relation between the individual's accomplishment age and his mental age, and use these terms to represent empirical measures instead of assuming that they measure exactly the underlying facts of capacity and achievement. Detailed criticisms have been made of the accomplishment ratio, both from the analytical and the statistical points of view, by Toops and Symonds and by Chapman.[1]

It would carry us beyond the limits of our space to discuss in detail all the questions which are raised concerning the achievement quotient or the accomplishment ratio. We may, however, comment upon a few of the most important implications and their practical significance.

The accomplishment ratio seems to imply, in the first place, that the intelligence test score gives a measure of native capacity which is independent of training, on the one hand, and that the educational test score gives an independent measure of achievement, on the other hand. This does not mean that intelligence and achievement are unrelated. On the contrary, it is usually assumed by those who use the

[1] Herbert A. Toops and P. M. Symonds, "What Shall We Expect of the A.Q.?" *Journal of Educational Psychology*, XIII (December, 1922), 513–28, XIV (January, 1923), 27–38.

See also J. Crosby Chapman, "The Unreliability of the Difference Between Intelligence and Educational Ratings," *Journal of Educational Psychology*, XIV (February, 1923), 103–08.

accomplishment ratio that they are so closely related that it is possible to make them correspond almost exactly. What is meant is that the particular tests which are used to measure intelligence are not affected by the accident of training, or of other mental traits than intelligence, and that the achievement tests will respond delicately to the changes in education or in effort, while the intelligence remains constant.

The distinction between what is measured by the two types of test is not as cleancut and definite as is implied in the assumption above mentioned. This means that in our reasoning about these measures and their relationship to one another we must treat them as rough, empirical measures, and not as highly refined measures of independent variables.

In the second place, the accomplishment ratio implies that, so far as the intellectual factor is concerned, achievement is based upon the same capacities as are involved in intelligence. It implies, furthermore, that achievement is also affected by other factors, such as interest, effort, and training. If these other factors, according to the hypothesis, are controlled and made equal in their effect upon the score, the achievement score will correspond completely to the intelligence score. This is an extreme example of the theory of general intelligence, which holds that every type of achievement is dependent upon the same kind of capacity.

A consequence of the assumption that achievement is always based upon the same kind of intellectual capacity is that if everybody did his best, and the factors of training and physical and mental environment were the same, the achievements of all individuals of the same mental age would be identical.

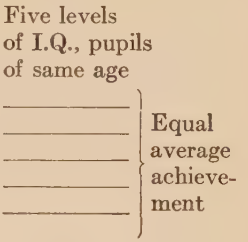
The prospect of being able to bring the accomplishment

of every individual into exact harmony with his potential achievement is a pleasing one to contemplate, but it probably cannot be done with anything like the exactness which is implied in using our present measures in the manner which has been indicated. The assumption that the intellectual factor in achievement in the various school subjects corresponds perfectly, and corresponds with general intelligence, is very doubtful, to say the least. A further assumption, which follows from this, that the variation which we find between the relation of the intelligence scores to the achievement scores is due solely to non-intellectual factors, such as effort, and so on, so that the individuals of a given mental age who make high achievement scores may be credited with an approach to maximum effort, whereas those who make low scores must be considered as making very deficient effort is, again, a doubtful assumption.

It is commonly observed that pupils of a given mental age whose I.Q.s. are low have a higher A.Q. than those whose I.Q.s. are superior. This is interpreted to mean that pupils of a low I.Q. work more nearly up to their capacity because they are stimulated more vigorously than their brighter companions. This greater stimulation of dull pupils very likely occurs, and it may account in part for the fact we are discussing, but it is certainly not the sole cause. Pupils of higher I.Q. must have a lower A.Q. as a consequence of the fact that the ability measured in the intelligence test and the ability required for achievement in the various subjects is not identical.

Take the extreme case in which there may be assumed to be no correlation between intelligence and achievement in a particular subject. We might take five groups of pupils all of the same age, representing five different levels in I.Q. or M.A. Because there is, by hypothesis, no correlation between intelligence and achievement, in this case, the

groups would be in descending order in A.Q., since the higher the I.Q. the lower would be the A.Q. This descending order would appear in some measure wherever the correlation between intelligence and capacity in the subjects is less than perfect. The assumption which is commonly made is that the correlation *would be perfect*, if it were not for variations in effort and other factors apart from general intelligence. It is safe to say that this assumption is false.



Douglass has discussed the same point and gives statistical proof that a negative correlation between the I.Q. and the E.Q. is due to "the unique nature of the correlation coefficient between a variable and a ratio of which the first variable is the denominator." [1] The correlation will be negative except when the correlation between I.Q. and E.Q. is 1, that is, when the intelligence quotient represents perfectly the ability to achieve. The assumption that this is so, as has already been said, is false.

The achievement quotient, then, is not a precise statistical measure and should not be used as an indication of the relation between the pupil's capacity and his achievement except in a very rough and approximate degree. In particular the contrast between the achievement quotient of pupils of lower intelligence and those of high intelligence has no educational significance. If a comparison between intelligence scores and achievement scores is made it should be only for the purpose of identifying cases of extreme dis-

[1] Harl R. Douglass and C. L. Huffaker, "Correlation between Intelligence Quotient and Accomplishment Quotient," *Journal of Applied Psychology*, XIII (1929), 76–80.

crepancy. This discrepancy may be made the starting point for further diagnosis and attempts to explain the discrepancy, and to find means of removing the cause when it is found.

To this difficulty is added one which grows out of the difference between the numerical significance of E.Q.s and I.Q.s. Data have been assembled by Rand [1] which show that the distribution of E.Q.s is narrower than is the distribution of I.Q.s. If this difference exists it necessarily causes the higher A.Q.s $\left(\frac{\text{E.Q.}}{\text{I.Q.}}\right)$ to be lower and the lower A.Q.s to be higher than are the corresponding I.Q.s.

7. *Norms*

The word *norm* may be used in two senses. In the first place, it may be taken to mean a standard of comparison to which it is implied that the various individuals of a group should conform. In the second place, it may mean simply the central tendency of the scores of a specified group without any implications concerning the desirability of individuals conforming to it. We shall use the term here in the neutral sense of the central tendency of a specified group.

The norm is inherent in the score of the age scale. This means that the standardization of the age scale, and the nature of the score in this scale, is of such a character that the relationship between the score of the individual and the average of the group is apparent in the score itself. We may pass immediately then to the discussion of norms in point scales.

The central tendency, which is taken as the norm, is ordinarily either the median or the arithmetical mean of the

[1] Gertrude Rand, *op. cit.*

group to which the norms apply. Norms may be classified according to the basis which is used for making up the groups.

The most widespread and significant norms are those which are based upon age grouping. Age norms consist of medians or means of the scores made by children of successive age groups. The chief problem which arises in the determination and interpretation of age norms is the selection of the cases. The selection must always be restricted to some degree. An age norm can never represent all the children of that age in the entire world. It is hardly likely, at the present time at least, that we can secure norms which represent the children even of the civilized world. The widest area which has been covered in any serious attempt to secure age norms has been a single country.

Within any one nation, such as the United States, there are numerous groups which may differ from one another in intellectual capacity. For example, there are environmental groups. Environmental groups would be represented by the children belonging to a particular neighborhood in a particular community. There are racial groups, sex groups, and occupational groups. If our aim is to secure norms which shall be representative of all the inhabitants of the country, and if we cannot test every individual of a given age, as we obviously cannot, it is necessary for a completely representative norm to secure a sample in which the individuals of each group were represented in the same proportion as they are represented in the population as a whole — assuming, of course, that there are differences in intelligence between these groups.

It may be said at once that no such systematic method of sampling, in order to secure age norms for children, has ever been carried out. Two chief methods for the purpose of securing an approximation through random sampling have

been employed. The first one, which was employed by Terman in his standardization of his first Stanford Revision of the Binet scale, was to select a community which might be presumed to represent neither extreme of ability, and then to test all of the children of that community. The other method, which is more commonly employed in securing norms for group point scales, is to test as large a number of children as possible, of various races and in various parts of the country, and then to assume that the different groups will be represented in the total group in the same proportion as they are in the country as a whole. These methods probably secure norms which fairly well approximate norms to be secured by a more systematic method of sampling.

The two chief methods which have been used to secure a sampling for standardization of tests which will be representative are to get a sample which will agree with the geographical distribution of the population as a whole, and to get one which will agree in occupational distribution. The selection of a sample to conform to the distribution of occupations has been used systematically for a number of years at the Institute of Child Welfare of the University of Minnesota. This method was used, for example, in the standardization of the Minnesota Pre-school tests. The distribution by occupations and also by geographical regions was taken into account by Terman and Merrill in the standardization of the second revision of the Binet scale.

Tests have practically all been standardized upon school children, and norms have been secured from children in the school. This results in the limitation of the norms to those ages at which practically all of the children are in school and can be tested. Satisfactory age norms have usually been limited to the ages seven to thirteen. Below age seven some children are not yet in school, and beyond age thirteen

some of the children, the brighter ones, of course, have gone on to high school and are not usually included in the testing program. If the tests are extended to the high school there begins to be an elimination of the duller pupils beyond the age of fourteen. Norms for the ages fifteen and above are therefore less representative than for the ages below.

In recent years the limitations on the samplings of pupils in school have been more clearly recognized than formerly and serious efforts have been made to secure adequate samples for the pre-school ages and for the adolescent and adult years. These attempts have been reasonably successful.

Age norms have been criticized on the ground of the alleged fact that stages in intellectual development are not well represented by chronological age. It is a well-established fact that individual children differ widely in the rate at which they mature physiologically, and that children of a given chronological age represent rather widely different stages of physiological maturity. It is believed, further, that the rate of intellectual maturing corresponds more closely to physiological maturity than it does to chronological age, and that if we can find a convenient measure of physiological maturity it is desirable to substitute an index of physiological maturity for chronological age in establishing norms and in comparing the scores of individuals with the norms.

It is not certain, however, that intellectual maturity corresponds more closely to physiological maturity than the chronological age, although it would seem natural that it should do so. In a study by T. M. Carter, the partial correlation, with age constant, was calculated between mental age and the ratio of ossification of the bones of the wrist. This ratio of ossification is the best measure that we have up

to the present of physiological maturity.[1] The correlation was found to be practically zero. If we take the mental age of an individual to be determined both by his intelligence and by the degree of his maturity, and if intelligence is not related to the rate of maturing, then mental age should be correlated with the measure of the stage of physiological maturity to the extent that the degree of mental maturity is represented by mental age, on the one hand, and is related to physiological maturity, on the other hand. Since we find no correlation we must conclude either that there is not a significant difference in the rate of intellectual maturing or that the rate of intellectual maturing does not correspond closely to the rate of physiological maturing. Carter found, further, that there was a closer correlation between mental age and chronological age than between mental age and ratio of ossification.

Abernethy made an extensive investigation of the relation between physical and mental growth and discovered a slight correlation between the various measures of physical growth and of intellectual growth.[2] The correlation was highest in early adolescence and at most was about .20 or .30 in a few specific instances.

8. Grade norms

Grade norms have been used less frequently with intelli-

[1] Thomas Milton Carter, "A Study of Radiographs of the Bones of the Wrist as a Means of Determining Anatomical Age." Unpublished Doctor's thesis, Department of Education, University of Chicago, 1923. See also Frank N. Freeman and Thomas M. Carter, "A New Measure of the Development of the Carpal Bones and Its Relation to Physical and Mental Development," *Journal of Educational Psychology*, XV (May, 1924), 257–70.

[2] Ethel Mary Abernethy, *Relationships between Mental and Physical Growth.* Monographs of the Society for Research in Child Development, Vol. I, No. 7. Washington: Society for Research in Child Development, National Research Council, 1936.

gence tests than with educational tests. Their interpretation is much more ambiguous than is the interpretation of age norms. This is due to the fact that the age composition of a grade in one school system may be very different from the age composition in another system. The amount of retardation or acceleration differs greatly from one community to another, and the age at entering school may also differ. If the pupils of a grade in one city have the same average intelligence scores as the pupils of the same grade in another city, it might be due to the fact that they possess the same intelligence, or it might be due to the fact that pupils in one community had a higher intelligence and were also farther advanced in the school. Furthermore, even with pupils in a given community, and with a given average intelligence, it might be possible to change the composition of a grade without changing the average intelligence score. That is, dull pupils might be eliminated from the grade and bright pupils added to it. To put it in another way, the composition of a grade is determined by the promotion policy which is in force.

If both age and grade norms are furnished with a test, and if the school administrator compares the scores made by the children of his system with both norms, he is likely to meet a situation which seems, at first sight at least, to be anomalous. He may find that the majority of the children are up to the age norms, but that a large majority make scores inferior to the grade norms. This has been found to be true, for example, in the use of the Haggerty scale, Delta 2. This situation is difficult to interpret. In order to interpret it, it is necessary that one have at his command all the facts concerning the grade progress of the children whose scores furnish the basis for the norms, and also of the children in the system which is being tested. These facts are never at hand. It seems, therefore, that grade norms for intelligence tests are of little practical value.

9. Norms for sex, race, and for social groups

In the discussion of age norms it was assumed that composite scores should be secured from all of the groups composing a community. It is sometimes held, however, that separate norms should be found for various groups, and that individuals of these groups should be judged each by comparison with norms for his particular group.

In considering the desirability of such group norms we must raise two questions. There is first the question of fact as to whether there exist sufficient differences between groups to make the norms desirable. If no significant differences between groups are found, then, of course, separate norms would have no meaning. If significant differences are found, then we are faced with a different question. What is the purpose of norms, and will this purpose be better served by differential group norms, or by composite norms? We may consider these questions individually with reference to the three types of group norms which have been proposed.

Sex norms. The prevailing view at the present time is that sex differences in intelligence tests are so small as to make it unnecessary to calculate separate norms for boys and girls. In fact, separate averages for the two sexes are not furnished with the majority of intelligence tests. Yerkes, Bridges, and Hardwick, in their original report on the point scale, stated that sex differences were large enough to demand separate norms. In the revised edition by Yerkes and Foster, however, the following statement is made: (p. 87) "On the basis of total score for the entire scale no significant sex differences can be made out from the original point scale results, but there seem to be sex differences in the ease with which certain of the individual tests are passed." In Terman's report upon the Stanford Revision of the Binet scale,[1]

[1] *The Stanford Revision and Extension of the Binet-Simon Scale*, Chap. IV.

he writes that there are slight differences in favor of the girls up to age thirteen. These differences, however, amount to only from two to four per cent and he does not consider them sufficient to warrant separate norms. Woodrow, from a study of a small group, concludes that the girls are superior to the boys, but that this superiority is not as great as we should expect from their comparative precocity in physiological development. He calculates, therefore, that girls in reality are inferior to the boys.[1] In view of the facts which were mentioned above concerning the relation between physiological maturity and mental maturity, this conclusion is, to say the least, a hazardous one. The best evidence which is now available indicates that sex differences in general intellectual capacity are negligible so far as the construction of norms is concerned.

Race norms. The existence of race differences in intellectual capacity will be discussed at some length in a later chapter. We may anticipate the conclusion of that discussion so far as to say that there appear to be significant differences between certain races in the capacity which is measured by our general intelligence tests. Whatever may be the ultimate explanation of these differences, they do now exist as a matter of objective fact. The largest differences of which we now have evidence are between the negroes and the Indians on the one hand, and the whites taken as a group, on the other hand.

Granting that these differences exist, does it follow that we should have separate norms? This raises the question concerning the purpose of the norms. Those who favor racial norms would say that the purpose of norms is to determine which individuals are normal and which, in distinction from them, are above normal by various degrees, or below normal by various degrees. They would add that

[1] *Brightness and Dullness in Children*, p. 121.

what is normal for one race is different from what is normal for another race. To apply to a race a standard which would result in rating a large majority of the individuals subnormal would be to contradict the meaning of normal. Normal, according to this view, is that which is usual, and therefore the majority of individuals of any group must be rated as normal.

If we base our decision upon this rather formal definition of the normal, we still have the alternative of considering a particular racial group as a distinct unit, or as a part of a composite group, which is made up of all the inhabitants of a given community. The treatment of a racial group as separate and distinct does not grow out of the necessities of the case, but must be justified by showing that such treatment gives ratings which are of greater practical usefulness than are obtained from composite ratings.

Take an illustration. The application to inferior racial groups of composite norms results in classifying a larger number as feeble-minded than would be so classified by the application of racial norms. Are the individuals who are thus classified as feeble-minded comparable to the smaller number of the superior race classified as feeble-minded, and do they demand the same treatment? The same question could be applied to the classification of the individuals at the upper end of the intellectual scale.

The prevailing view would probably be that so far as measurement by any absolute standard is concerned, the larger number of the inferior group which is rated feeble-minded by a test is comparable to the smaller number of the superior group which is so rated. Feeble-mindedness, in other words, represents a certain *degree* of capacity, and not the existence of a trait which is absent from normal individuals, nor the absence of a trait which is present in normal individuals. At any rate, this is true of degrees of deficiency above that of feeble-mindedness.

On the other hand, the kind of rating which should be given an individual depends upon the significance which that rating has with reference to his capacity to adjust himself in his environment. If an individual of an inferior race intellectually comes into contact and competition wholly or chiefly with other individuals of the same race, successful adjustment demands a lesser degree of ability than if he comes into contact and competition with individuals of a superior race. If the individuals of a race, then, are largely segregated in their social, industrial, and commercial life, it would seem preferable to apply to them norms which have been derived from their own group. If, however, they are mixed with individuals of another race in their social and vocational activities, they should be rated by composite norms. It may be that we should apply composite norms when tests are used for some purposes and separate norms when they are used for other purposes. In any case the issue is one which should be decided pragmatically rather than on the grounds of a formal definition of normality.

Even if we grant the desirability of having race norms for certain purposes, there are two difficulties in the way of securing usable norms. The first difficulty arises out of the fact that the standing in mental tests is affected by the social environment of the individual as well as by his race. Segregated race groups differ in social environment. Their scores are therefore due to the compound result of race and environment and it is impossible to disentangle the share which is contributed by race from that which is contributed by surroundings. The only method by which an approximation to comparable race norms may be secured is to obtain scores from the different races which live in the same social environment. This will enable us to eliminate differences which are due to gross external circumstances. They

may not, however, eliminate differences due to general cultural background.

The second difficulty is that of racial mixture. A racial norm would apply only to those of pure blood. In an experiment with the army test,[1] the scores of a group of mulattoes of lighter skin were compared with another group of darker skin. In the Army Alpha the lighter-skinned group made a median score of 50 and the darker-skinned group a median score of 30. Furthermore, the percentage of darker negroes was greater among the illiterates than among the literates. Garth found similarly that Indians of mixed blood made higher scores than Indians of pure blood. These facts indicate that racial norms which are adapted to those of pure blood would not apply to those of mixed blood. Since it would be very difficult either to secure norms for those of mixed blood or to determine the degrees of mixture in the case of individuals, the application of norms to those of different races becomes one of large practical difficulty.

Social norms. There are unquestionably large differences between the average scores of various social groups. This is true whether we compare those who live in various neighborhoods in the city, or whether we compare the city with the small town or the rural district, or whether we compare different sections of the entire country. Some have contended that the existence of these differences demands norms for social groups. This involves questions similar to those which are raised in discussing racial norms.

In the first place, we must consider the purpose for which norms are created. They constitute standards by which individuals may be compared with one another through the medium of the standard. The question whether we should have norms for different social groups reduces itself then, to this question, Do we wish to compare directly only individ-

[1] R. Yerkes (Ed.), *Psychological Examining in the U.S. Army*, p. 735.

uals of a given social group, or do we wish to compare them directly with individuals of another group? The individuals of the various social groups do come into competition with one another to a greater extent than do individuals of different races, at least in the case of the negroes and the Indians. The assumption of our social organization is that the opportunity for free intermingling and competition is a complete one. If this is the case it would seem to lead to the conclusion that we ought to have norms which can be applied alike to all.

In the case of social level norms, however, there is another question involved. The differences between the various social groups may be held to be due not to inherent differences, but to accidental differences of training and environment. In so far as this is the case it may legitimately be held that we cannot get at the individual's real native capacity by his raw score. We must make allowance for his training.

The use of social norms, however, would ascribe the entire difference between social groups to the differences of their environment. Most psychologists would regard this allowance as too great. They would hold that segregation into a social group is to some extent a selective process based upon intelligence, and that, therefore, there is a real native difference between such groups. They would hold furthermore, that there is an interaction between the effect of native capacity and environment. The poor environment is unfavorable to the development of native capacity, while, on the other hand, intelligence more or less creates its own environment by the fact that individuals of meager intellect allow their environment to deteriorate, while those of higher capacity improve their environment. The two factors are therefore so entangled that it becomes almost impossible to determine how much each is responsible for the group differences which we find. Environmental norms, therefore, would be based

in part upon an error in assumption, and it would be impossible to determine how great that error was.

The construction of social norms is attended with the further difficulty, analogous to that which was mentioned in connection with race norms. It would be exceedingly difficult to find a method of grading social environments so as to apply norms to them. Furthermore, many gradations of environment could be found and the same individual is subject to the influence of more than one environment. For example, his home environment may be of one sort and his school environment of another. These complications and difficulties seem to make it inadvisable to create norms for social groups.

10. The use of local norms

Because of the difficulties and complications in the interpretation of norms which are based upon large-scale testing, it is frequently more serviceable to use the average of the group which is being tested as a provisional norm, rather than to use the general norms which have been derived for the general use of the test. The greater number of practical uses to which tests are put demands simply that individuals of a group be rated in comparison with one another. This comparative rating is more easily done if the norm which is used agrees with the average standing of the group. This is not likely to occur when general norms are used. The desirability of using local averages rather than general norms is particularly great in those cases in which the general norms have not been established on a large number of cases selected at random. Only in the case of the most thoroughly and carefully standardized test are the general norms to be relied upon.

In cases in which an individual's general intelligence is to be estimated, one may use such general norms. It is usually

advisable, however, to base such a rating upon the scores of several tests rather than of one test alone. The variations which are frequently found in the intelligence rating of an individual by different scales makes this precaution necessary. When it is desired only to get a comparative rating of the individuals of a group for classification or for other similar purposes, the use of local norms is to be advised.

Chapter XII

HOW TO TABULATE THE RESULTS OF TESTS

THE purpose of the present chapter is to indicate and illustrate the steps which one should go through in tabulating the results of tests, so that one may arrive at their interpretation. The purpose is not to make the reader familiar with statistical methods. It is not to describe how one proceeds in calculating an average or a median, or a probable error, or a coefficient of correlation. For information concerning these matters the reader is referred to books on statistics. The information here given supplements that which is given in books on statistics, but is not a substitute for it.

1. Tabulating the scores

In choosing an example to illustrate the steps to be taken in tabulating the scores, a class the size of the ordinary public school class is taken as a unit. The group which is used consists of fifty pupils. A larger group would, of course, give measures which are statistically more reliable. The individual teacher, however, frequently has occasion to tabulate the scores of a single class, and the procedure which is appropriate for a larger group is not always appropriate for a group of this size. On the other hand, everything which may be legitimately done with a group of fifty may also be done with a larger group. Furthermore, the statistical measures — the median, the quartile deviation, the correlation coefficient — are reliable enough with a group of this size to have practical meaning.

In order to show each step from the beginning, we shall start out with the original table of scores. Table XIII shows

the scores of each child upon all of the tests which are to be brought into the comparison. In this particular case we have the age, the I.Q. on the Stanford Revision of the Binet test, the I.Q. on the Otis test, the Otis score, the score on the Haggerty test Delta 2, the score on the Gray Oral Reading test, the score on the Burgess Silent Reading test, and the score on an arithmetic test. By including certain subject-matter or educational tests we shall be able to show applications of intelligence tests that we could not otherwise illustrate. The numbers in the first column of the table represent the individual children of the class. Each horizontal row of scores, then, was made by one particular child.

It is obvious that we cannot make general comparisons or draw conclusions from large numbers of individual scores when they are merely tabulated in this form. They are too numerous for us to summarize them by inspection. It is therefore necessary to calculate summary scores. We have therefore found the average score of the girls as a group, of the boys as a group, and of the entire class, for each of the tests. Thus we see that the average I.Q. of the girls is 117.1, of the boys, 124.15, and of the entire class, 118.9. In this particular class the boys have a higher I.Q. than the girls.

It will be noticed that the average for the entire class is not the average of the two averages for the boys and girls separately. The reason for this is that there are more girls than there are boys, and therefore the scores of the girls as a group have greater weight in the average of the entire class than do the scores of the boys. In order to find the average of the entire group, it is necessary to take the total score for the whole group and divide by the total number of cases. Only in the case that the two sub-groups have the same number of cases is it legitimate to average their averages.

Now that our attention has been drawn to the difference

TABLE XIII. INDIVIDUAL SCORES OF A GROUP OF FIFTY CHILDREN ON A NUMBER OF TESTS

GIRLS

	AGE IN 1923	I.Q. BINET	I.Q. OTIS	OTIS SCORE	HAGGERTY SCORE	GRAY SCORE	BURGESS SCORE	ARITHMETIC SCORE
1.	11	101	119	48	125	61.25	100	
2.	11	119	111	43	87		68	58
3.	11	117	107	43	120	56.25	80	48
4.	11	106	115	45	91	58.85	92	66
5.	11	120	117	50	100	57.5	80	52
6.	11	109	126	60	125	58.75	92	64
7.	11	122	103	32	125	63.75	92	54
8.	11	119	120	51	97	53.75	98	58
9.	11	130	121	49	103	53.75	50	64
10.	11	113	126	59	127	63.75	98	56
11.	10	113	120	46	126	62.5	86	66
12.	11	105			95	52.5	56	50
13.	10	133	127	53	119	60	92	68
14.	10	143	129	56	110	53.75	62	66
15.	10	125	125	47	103	57.5	62	62
16.	11	130	121	49	126	56.25	68	62
17.	11	118	130	59	127	62.5	86	64
18.	12	108	117	55	128	60	86	60
19.	11	114	110	40	103	57.5	62	52
20.	10	123	117	42	106	60	74	54
21.	11	116	113	44	123	68.75	74	66
22.	10	126	120	47	103	62.5	68	68
23.	11	109	116	45	126	62.5	80	54
24.	11	107	109	42	118	53.75	68	60
25.	11	116	117	51	114	57.5	80	60
26.	10	126	125	47	123	60	86	60
27.	10	136	125	50	118	75	80	58
28.	12	100	90	31	90	52.5	50	
29.	11	117			149	61.25	100	68
30.	12	106	92	29				
31.	11	119	124	54	118	53.75	68	62
32.	11	104	105	41	104	55	62	68
33.	10	128	120	47	141	60	74	60
34.	10	112	111	36	109	63.75	92	58
35.	10	112			111	63.75	80	56
36.	11	117			114	67.5	92	
37.	11	112	122	51	120	65	98	68
Total		4331	3850	1542	4124	2092.5	2836	1990
Average....		117. 1	116.66	46.72	114.55	59.78	78.78	60.6

TABLE XIII. INDIVIDUAL SCORES OF A GROUP OF FIFTY CHILDREN IN A NUMBER OF TESTS (*continued*)

BOYS

	AGE IN 1923	I.Q. BINET	I.Q. OTIS	OTIS SCORE	HAGGERTY SCORE	GRAY SCORE	BURGESS SCORE	ARITHMETIC SCORE
38.	10	126	132	60	113	57.5	68	62
39.	10	123	126	52	117	53.75	74	56
40.	12	139	118	55	133		80	68
41.	11	121	118	48	124	50	86	62
42.	9	147	125	46	132	70	100	64
43.	11	112	118	47	93	55	62	58
44.	11	137	120	58	135	52.5	56	
45.	10	114	113	39	132	57.5	80	70
46.	10	127	111	37	122	57.5	62	62
47.	11	114			119	62.5	80	
48.	13	124	107	52	149	51.25	56	
49.	11	116			123	53.75	100	64
50.	11	114	110	43	115	60	92	56
Total.....		1614	1298	537	1607	681.25	996	622
Average....		124.15	118	48.82	123.61	56.77	76.61	62.2
Total Girls.		4331	3850	1542	4124	2092.5	2836	1990
Total Boys		1614	1298	537	1607	681.25	996	622
Grand Total....		5945	5148	2079	5731	2773.75	3832	2612
Average....		118.9	117	47.25	116.95	59.01	78.20	60.74

between the boys and girls, let us examine the other averages, and see whether they give results consistent with the averages of the I.Q.s. The Otis I.Q. again gives higher averages for the boys, as does the Otis score. It does not follow, of course, that because the boys have a superior average Otis score, they should also have a superior average Otis I.Q. If they were sufficiently older than the girls the average score might be higher and the I.Q. lower. Again, we find that the boys make a higher score on the Haggerty

test. Thus, on all the intelligence tests the score of the boys is above that of the girls. When we examine the educational tests, on the other hand, the girls are superior.

The explanation of the difference between boys and girls in the relation between capacity and achievement is not the chief problem before us. It may be said that this is not an isolated finding, and that the chief suggestion toward an explanation is that either the school work is more suited to the interest of the girls than of the boys, or that girls are more conscientious and studious.

While the average gives us a number which is readily grasped and which is convenient for making comparisons, it does not tell us all that we need to know about the scores of a group. The entire list of individual scores is too complex to grasp, but, on the other hand, the single summary figure which is represented in the average is too much simplified to give us all the information we need. It does not tell us, for example, whether the scores cover a wide range or a narrow range, or whether the largest number of scores fall in one part of the range or in another part.

2. The distribution table

In order to secure more information about the distribution of the scores than is given in the average we may tabulate them in such a way as to show the number of individuals who make the various scores. A table in which the scores are classified in this way is called a distribution table. The distributions of the scores given in our basic table are shown in Table XIV and Table XV. The meaning of these tables may be gained from the distribution of the Binet I.Q.s shown in Table XIV. This distribution shows that there was one child whose I.Q. was in the class 145–149, another whose score was in the class from 140–144, three whose I.Q.s were in the class 135–139, and so on.

Table XIV. Distribution of I.Q.s

Class Intervals	Binet Scale	Otis Scale
	No.	No.
145–149	1	
140–144	1	
135–139	3	
130–134	3	2
125–129	6	9
120–124	6	9
115–119	10	10
110–114	10	7
105–109	7	4
100–104	3	1
95 –99		
90 –94		2
Total	50	44
Median	117.5	119

Let us go back for a moment and consider the steps which are gone through in constructing such a distribution table. It will be noticed that the scores are classified or grouped. The table does not show how many individuals make each particular score, but rather how many make scores which

TABLE XV. DISTRIBUTION OF THE SCORES IN FIVE TESTS

OTIS TEST		HAGGERTY TEST		GRAY READING TEST		BURGESS READING TEST		ARITHMETIC TEST	
Class Intervals	No.	Class Intervals	No.	Class Intervals	No.	Class Intervals	No.	Class Intervals	No.
60–64	2	145–149	2	74–76	1	100–104	4	69–71	1
55–59	6	140–144	1	71–73		95– 99	3	66–68	10
50–54	9	135–139	1	68–70	2	90– 94	7	63–65	5
45–49	13	130–134	3	65–67	2	85– 89	5	60–62	11
40–44	8	125–129	9	62–64	9	80– 84	9	57–59	5
35–39	3	120–124	7	59–61	8	75– 79		54–56	7
30–34	2	115–119	7	56–58	11	70– 74	4	51–53	2
25–29	1	110–114	5	53–55	9	65– 69	6	48–50	2
		105–109	2	50–52	5	60– 64	6		
		100–104	6			55– 59	3		
		95– 99	2			50– 54	2		
		90– 94	3						
		85– 89	1						
Total	44		49		47		49		43
Median	48.7		118.9		58.6		81.9		61.5

fall within a particular class. The range of scores which was used in this table in making up the classes is five points. If the frequency of each individual score had been tabulated the scattering would have been too great to enable us to determine where the greatest frequency lies. We have therefore grouped the scores into classes and have chosen

class intervals that will give from ten to fifteen classes or groups. With a small number of cases the number of class intervals should be fewer than with a large number of cases.

The first step in making a distribution table, then, is to determine what our class intervals shall be. We may do this by finding the highest and the lowest score and the difference between them. This will give us the entire range of the scores. We may then provisionally divide this entire range by ten, or some larger number if we have a large number of cases. This will give us the range of each class interval. For example, in the case of the Binet I.Q.s the lowest score is 100 and the highest 147. The difference, 47, divided by 10 gives us 4.7. Since 5 is the nearest whole number to 4.7, and since 5 is a convenient class interval, we select this for our range.

The next step is the very simple one of finding the number of scores in each class interval by the method of tallying. The form of the tally record is illustrated in the second column of Fig. 17, page 333. By adding the tallies of each class we have our distribution table.

From the distribution table it is easy to calculate the median. The medians in most cases are not very different from the arithmetic mean, which was the average used in the basic table. Since the median is easy to calculate, it may be used in place of the arithmetic mean where we wish to obtain an approximate average for the distribution.

We may learn the characteristics of a distribution by inspecting the distribution table. They are brought out more clearly, however, by means of a chart. The most commonly used type of graph of a distribution is the column diagram, or histogram. The histogram of the Binet I.Q.s shown in Table XIV is given in Fig. 16.

This histogram shows at a glance that the distribution is skewed. The upper part of the distribution conforms fairly

closely to what is called the normal distribution frequency. The lower part, however, has the appearance of being cut off rather abruptly. There are no I.Q.s below 100. In an

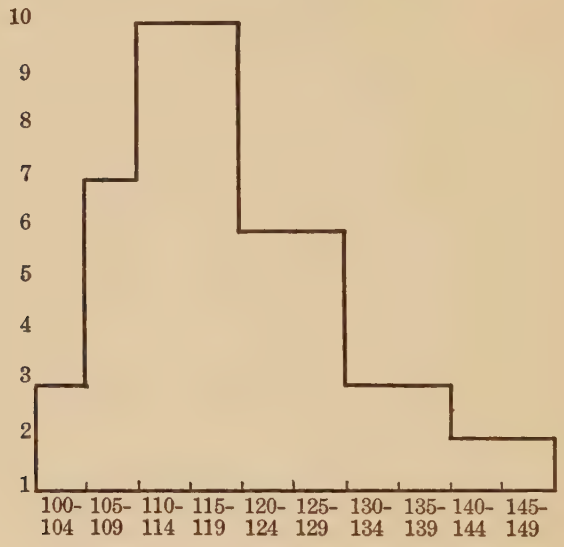


Fig. 16. Histogram of Binet I.Q.s

unselected group, of course, there would be as many below as above 100. It would appear, therefore, that the children of this class are not representative of the population as a whole, but represent only the upper half of the population. We need not here discuss the question whether the evident superiority of these pupils is due wholly to innate capacity, or in part to training.

The form of the distribution throws light upon two matters, first, the selection of the cases, and second, the suitability of the test. Take first the selection of the cases. Numerous distributions of scores have convinced psychologists

that abilities are distributed, at least approximately, according to the normal frequency distribution, provided there is a random sampling of individuals. We have a random sampling when the frequency of the cases representing different degrees of ability are the same as the frequency of corresponding abilities in the population as a whole. The most prominent characteristic of the normal distribution is that the largest number of cases occur in the middle and that the two sides of the curve of distribution are symmetrical.

If the sampling is not random, the curve of distribution is likely to be unsymmetrical. An unsymmetrical curve, however, may also be an indication of the inadequacy of the test. If the test is too hard for the group, the largest number of scores will fall toward the bottom of the scale, and the distribution curve will be skewed toward the upper part of the scale. If the test is too easy, the largest number will fall toward the upper part of the scale and the curve will be skewed toward the bottom. We cannot be sure, therefore, from the form of the curve, what the cause of the skewness is. A skewed curve should lead us to pursue our investigation until we arrive at its explanation.

3. The percentile curve

Another useful form of graphic representation of a distribution is a percentile curve. The advantages of a percentile curve are thus set forth by Otis.[1] "A percentile curve shows at a glance not only the median score of a class, but also the range and variability of the scores. It shows at a glance just what per cent of the scores of a class is exceeded by the score of any given individual, and just what per cent of the class attains or exceeds any given score. Two or more curves on the same graph show very vividly the amount of overlapping of the scores of different classes."

[1] Arthur S. Otis, *Otis Self-Administering Tests of Mental Ability, Manual of Directions*, p. 10. Yonkers-on-Hudson, N.Y.: World Book Co., 1922.

To get an understanding of the percentile curve, let us go through the procedure by which it is made up. (See Fig. 17.) We start from the distribution table as before, and in this illustration we use the distribution of the Binet I.Q.s. The first column at the left of the chart shows the class intervals of the scores. The next column shows the first step in making a distribution table. It contains the tallies of the scores which fall within the various class intervals. In the next column we depart from the distribution table. Instead of writing in the number of cases in each class interval, we write in each space the total number of cases in that class interval plus all of those in the lower intervals. The figures then represent the cumulative frequency. In the next column these cumulative frequencies are translated into terms of the percentage of the total number of cases.

We are now ready to construct the graph. The scale along the bottom of the chart represents the percentage of cases. The vertical scale constructed in the middle of the chart represents the scores. The significance of the curve in general is this. Each point on the curve represents the percentage of the group which makes a given score or lower. Thus, in this particular case, 20 per cent of the children make a score of 110, or lower; 40 per cent make 115 or lower; and 90 per cent make a score of 135 or lower.

Before commenting further upon the facts which are shown by the chart, let us go back for a moment and trace the steps in constructing the curve. The procedure in brief is as follows. Place a point at the lower limit of the first class interval and at the zero point on the horizontal scale. Second, place a point at the upper limit of the first class interval and at the place on the horizontal scale which represents the percentage of cases in this interval. Third, place a point at the upper border of the second class interval, at the place on the horizontal scale representing the percentage

Score	Tally	Sub-Total	Per Cent
145-149	I	50	100
140-144	I	49	98
135-139	III	48	96
130-134	III	45	90
125-129	##I	42	84
120-124	##I	36	72
115-119	## ##	30	60
110-114	## ##	20	40
105-109	##II	10	20
100-104	III	3	6

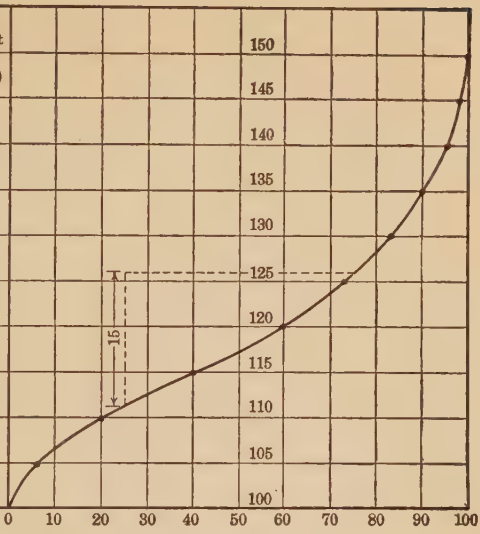


FIG. 17. PERCENTILE CURVE OF BINET I.Q.s

in the second group. Do the same for each of the other class intervals. The final step is to draw a smooth curve through the points which have been plotted.

The use of the percentile curve may be further illustrated. The median, of course, is the fifty percentile. This is found by noting on the vertical scale where the percentile curve crosses the 50 per cent line. It will be seen that this is 117.5. It agrees with our calculation of the median. An important measure of the characteristic of a distribution is the measure of its variability. A commonly used measure of variability is very easily found from the percentile curve. This is Q, which represents half the difference between the 75 percentile and the 25 percentile score, or the difference between Q_3, and Q_1 . Q_1 is found by locating the point on the curve above 25 on the horizontal scale, and Q_3 by locating

the point above 75 on the horizontal scale. These are indicated by the short cross-lines on the curve. In the present case, Q_1 is 111.3, Q_3 is 126, and $Q = \frac{126-111.3}{2}$ or 7.4. Twice 7.4 or 14.7 is the range which includes the middle half of the scores of the group. This is a convenient measure by which to compare various distributions with one another.

We cannot, of course, compare directly the semi-interquartile range or Q in order to get the relative variability of two distributions, unless the same tests were used in the two cases. For the procedure to be followed when the scale or test is different, the reader is referred to a book on statistics.

The percentile graph is useful, finally, as a means of giving a convenient measure of the position of the individual in the distribution and of comparing the individual's position in different distributions. Suppose that, in the present case, a pupil's I.Q. was 115. This would mean that his percentile rank was 40. In other words his I.Q. exceeds that of the lower 40 per cent of the group and is exceeded by the upper 60 per cent. If now, we wish to determine whether the pupil's score in arithmetic is higher or lower, relatively to that of the entire class, than is his I.Q., we can do so by finding his percentile rank in arithmetic and comparing it with his percentile rank in I.Q.

4. Correlation

The comparison just suggested between a pupil's percentile rank in two tests is a rough means of finding the correlation in an individual case. The next procedure is to tabulate the scores on pairs of tests, so as to bring out the relationship between the scores for the group as a whole. Take as an illustration the correlation table representing the relation between the Binet I.Q. and the Otis I.Q., Table XVI.

TABLE XVI. CORRELATION TABLE SHOWING THE RELATION BETWEEN BINET I.Q.S AND OTIS I.Q.S

BINET SCALE	OTIS SCALE									
	90–94	95–99	100–104	105–109	110–114	115–119	120–124	125–129	130–134	TOTAL
145–149								1		1
140–144								1		1
135–139						1	1	1		3
130–134							2	1		3
125–129					1		2	2	1	6
120–124			1	1		3		1		6
115–119				1	2	1	2		1	7
110–114					4	1	2	1		8
105–109	1			1		3		1		6
100–104	1			1		1				3
Total...	2	0	1	4	7	10	9	9	2	44

$r = .50 \pm .076$

A correlation table is really a simultaneous distribution of the scores of one test in one dimension and scores of the other test in the other dimension. In this particular case, the distribution of scores in the Binet test is represented vertically and the distribution of the scores in the Otis test horizontally. The total distribution of Binet I.Q.s is shown in the last column to the right and the distribution of the Otis I.Q.s in the lowest horizontal row.

Consider the make-up of the table from the point of view of individual cases. In the lower left-hand corner is the figure 1. This means that one child had an I.Q. on the Binet scale within the range 100–104, and an I.Q. on the Otis

scale within the range 90–94. Directly above him is represented a child whose Binet I.Q. is in the class 105–109 and whose Otis I.Q. is in the class 90–94. In the upper horizontal row, we find a tally representing a child whose Binet I.Q. is in the class 145–149, and whose Otis I.Q. is in the class 125–129. In these cases, the I.Q.s in the two tests correspond fairly well. A low I.Q. in the one test goes with the low I.Q. in the other, or a high I.Q. in the one test goes with a high I.Q. in the other.

In other cases, however, the correspondence is not so close. For example, one child has a Binet I.Q. which places him in the lowest section of the scale, but an Otis I.Q. in the class 115–119. It is easy to locate roughly those cases in which the scores on the two tests correspond and those on which they differ. All the cases which cluster about a diagonal line running from the lower left-hand to the upper right-hand corner are cases in which the two scores correspond. Those which fall in the upper left-hand or the lower right-hand corner of the table are cases in which there is a discrepancy. In the present comparison there are no cases of children who have a high I.Q. in the Binet test and a low I.Q. in the Otis, but there are a number of cases of children who have a comparatively high I.Q. in the Otis, but a low I.Q. in the Binet. It appears that the qualities which enable a child to do well in the Binet test also enable him to do well in the Otis test, but that there are certain qualities which make possible a comparatively high score on the Otis test, but which are not adequate to give a high score on the Binet test. What these qualities are would require further analysis. It may be that, since a group test is more largely a measure of speed than an individual test, rapidity of performance is the quality which gives a high score in the Otis test, but not in the Binet. This question illustrates the way in which a correlation table may be used

to make a further analysis of the scores than can be made from the distribution of the scores in the individual tests alone.

Another way in which a correlation table showing the distribution of the scores in two intelligence tests may be used is to discover the cases of children on whom one or the other of the tests appears to be unreliable. It is quite possible that, in any particular test, a child may do himself injustice due to the circumstances of the moment. If we find that a child makes a low score in one test and a high score in another, we should follow the matter up by giving him a third test, in this way determining more nearly what his true rank is. It is a common practice to give an individual test to a child when his scores on the two group tests show wide discrepancy.

For practical administrative uses the detailed correlation table gives most of the information on correlation which we need. In order to determine the amount of correlation between two tests so that it may be compared with the amount of correlation between other tests, it is necessary to express the correlation in terms of a single coefficient. This coefficient is derived by the use of one of the formulæ which are now available and which can be found in books on statistics. It will be remembered that the range of the coefficients is from -1, which expresses complete negative correlation, through 0, which expresses no correlation whatever, to $+1$ which expresses perfect positive correlation. The correlation coefficient has been calculated from Table XVI, and found to be $.50 \pm .076$, as stated at the bottom of the table. This is commonly regarded as a rather low correlation between intelligence tests. We usually expect the correlation coefficient between tests of the same nature to be .70 or higher. A possible explanation of the low correlation in this case is that the range of the distribution of I.Q.s

is narrow. In the case of the Binet test there is no I.Q. below 100. It will be remembered that the distribution is a skewed one, suggesting that the lower part has been cut off by some process of selection. Furthermore, the I.Q.s on the Otis tests are also nearly all above 100. This means that the pupils of this class are more nearly homogeneous in intelligence than are the pupils of an unselected group. The correlations between tests of a homogeneous group are always lower than the correlations in the case of a group of more widely scattered abilities.

Table XVII shows the correlation between the Otis score and the Haggerty score. This correlation is still lower than that between the Binet I.Q. and the Otis I.Q. We might expect that there would be a higher correlation between two group tests than between a group test and an individual

TABLE XVII. CORRELATION TABLE SHOWING THE RELATION BETWEEN OTIS SCORES AND HAGGERTY SCORES

OTIS SCALE	HAGGERTY SCALE													
	85–89	90–94	95–99	100–104	105–109	110–114	115–119	120–124	125–129	130–134	135–139	140–144	145–149	TOTAL
60–64						1			1					2
55–59						1			3	1	1			6
50–54			1	1		1	4	1					1	9
45–49		2		3				2	4	1		1		13
40–44	1			2	1		2	2						8
35–39					1			1		1				3
30–34		1							1					2
Total	1	3	1	6	2	3	6	6	9	3	1	1	1	43

$r = .302 \quad .093$

test. No explanation is suggested for this lower score, unless it be that the Haggerty test is less reliable in this particular case than is the Otis test. It is more probable, however, that the difference is an accidental one.

It will be noticed that, in the preceding tables, the I.Q.s were correlated with I.Q.s and the raw scores with the raw scores. It is not legitimate, in general, to correlate I.Q.s with raw scores or with mental ages. The only case in which this would be legitimate is the one in which all the pupils are the same age. In such a case, of course, the I.Q. and the mental age are comparable. The reason that we cannot correlate the I.Q. with mental age or with raw score is that I.Q. expresses the relative standing or brightness of the pupil, and remains the same from age to age, whereas mental age or raw score represents the attainment of the pupil on fixed scale. Thus, suppose that a very bright child was in a class with other children who were older than himself. He would stand high in I.Q. but would stand low or at least have a medium rank in mental age. There would thus appear to be a discrepancy between his intelligence quotient and his mental age. This discrepancy would not appear if he were ranked in terms of I.Q. in two tests, since in both cases his standing would be high. A discrepancy would not appear, furthermore, if he were ranked in mental age in two tests, since in this case his rank would be medium or low in both cases. The general rule, then, is that we should always correlate a relative score with a relative score, or an absolute score with an absolute score, but never a relative with an absolute score.

The next tables show the correlation between intelligence tests and educational tests. Table XVIII shows the correlation between the Otis score and the score in the Gray Oral Reading Test. It appears both from the inspection of the table and from the correlation coefficient that there is no

TABLE XVIII. CORRELATION TABLE SHOWING THE RELATION BETWEEN OTIS SCORES AND GRAY SCORES

OTIS SCALE	GRAY SCALE									
	50–52	53–55	56–58	59–61	62–64	65–67	68–70	71–73	74–76	TOTAL
60–64			2							2
55–59	1	1		1	2					5
50–54	1	3	2	1		1			1	9
45–49	1	2	3	3	3		1			13
40–44		2	2	2			1			7
35–39			2		1					3
30–34	1				1					2
Total	4	8	11	7	7	1	2		1	41

$r = -.07 \pm .105$

correlation between these two measures. The pupil's ability in oral reading does not seem to be determined by his brightness or his intelligence. This lack of correlation is, of course, to be interpreted in the light of the fact that we have a relatively homogeneous group. If the group contained pupils of very low intelligence, we should undoubtedly find their reading attainment to be also comparatively low.

It is rather more surprising to find evidence in Table XIX that there is little or no correlation between the intelligence test score and the score in the Burgess Silent Reading Test. We ordinarily expect a positive correlation between a measure of intelligence and a measure of silent reading. A moment's reflection, however, reminds us that the Burgess test is a rather easy one, and that it possibly does not discriminate satisfactorily between the silent reading abilities of pupils in the fifth and sixth grade. The

TABLE XIX. CORRELATION TABLE SHOWING THE RELATION BETWEEN OTIS SCORES AND BURGESS SCORES

OTIS SCALE	BURGESS SCALE											
	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100–104	TOTAL
60–64				1								2
55–59		1	1				1	2		1		6
50–54		1		1	1		3		1	2		9
45–49	1		2	2	1		1	3	1		2	13
40–44			2	2	2		1		1			8
35–39			1				1		1			3
30–34	1								1			2
Total	2	2	6	6	4	0	7	5	6	3	2	43

rather high correlation between the oral reading score and the silent reading score, as shown in Table XX, suggests further that, for these grades, the Burgess tests measure the more mechanical aspects of reading rather than the ability to get thought from the printed page.

Since there is practically no correlation between intelligence tests and these particular reading tests, we cannot use such a table as No. XIX to analyze and interpret the relative scores of individual pupils. The degree of correlation does not give us ground to expect that a high intelligence score will be accompanied by a high achievement in the subject. We are not justified in expecting that a particular pupil, because he makes a high score in the intelligence test, should make a high score in the subject-matter test. We cannot use the intelligence test in such a case to segregate pupils

TABLE XX. CORRELATION TABLE SHOWING THE CORRELATION BETWEEN GRAY SCORES AND BURGESS SCORES

BURGESS SCALE	GRAY SCALE									
	50–52	53–55	56–58	59–61	62–64	65–67	68–70	71–73	74–76	TOTAL
100–104		1		2			1			4
95– 99		1			1	1				3
90– 94			2	2	2	1				7
85– 89	1			2	2					5
80– 84			4		3				1	8
75– 79										
70– 74		1		2			1			4
65– 69		2	2		1					5
60– 64		3	3							6
55– 59	3									3
50– 54	1	1								2
Total	5	9	11	8	9	2	2		1	47

$r = .53 \pm .072$

according to their capacity, nor can we regard a high intelligence and a low achievement score as evidence of a lack of application. It is only when there is a fairly high correlation in general between the intelligence test and the subject-matter test that we can make such an administrative use of the scores.

In those cases in which there is in general a rather high correlation between the scores on two tests, it is appropriate to examine those cases which exhibit very wide discrepancy. In the case of the correlation between the Gray scores and the Burgess scores, for example, we find three children who make high scores on the Burgess test and low scores on the

Gray test. These children evidently have some specialized difficulty in oral reading. The detailed discussion of such a problem as this does not belong in a book on mental tests. The case is an illustration, however, of the sort of examination which may be made of the relation between an intelligence test and an educational test.

Our final table, XXI, gives the correlation between the Otis scores and the arithmetic scores. The correlation between these scores is low, but perhaps high enough to warrant the study of cases showing a very wide discrepancy. The pupil who makes a score of 70 on the arithmetic test and only 35 on the Otis test appears either to have specialized ability in arithmetic, to be very industrious, or to be incorrectly rated on the Otis test. The examination of other evidence concerning the child's ability would probably indicate which of these suppositions is the correct one.

The aim of the foregoing discussion has been to indicate briefly some of the chief ways in which the scores of mental

TABLE XXI. CORRELATION TABLE SHOWING THE RELATION BETWEEN THE OTIS SCORES AND THE ARITHMETIC SCORES

OTIS SCALE	ARITHMETIC TEST								
	48–50	51–53	54–56	57–59	60–62	63–65	66–68	69–71	TOTAL
60–64					1	1			2
55–59			1		1	1	2		5
50–54		1	1	2	2		2		8
45–49			1	1	5	2	3		12
40–44	1	1	2	1	1		2		8
35–39				1	1			1	3
30–34			1						1
Total	1	2	6	5	11	4	9	1	39

tests may be tabulated and used to promote the efficiency of teaching and the handling of pupils. The reader will have been impressed with the fact that the results of tests should be used cautiously, and that a hasty application of them should be avoided. Teachers should be on the lookout for discrepancies, and should attempt to follow them up in order to arrive at their correct interpretation. Through the repeated use of tests and the analysis of their results, the teacher and the principal or the supervisor should gradually gain a notion of the general capacities, the special capacities, and the weaknesses of individual pupils.

The account in this chapter has been designed to serve the teacher or principal, rather than the research officer or the superintendent. The illustrations have purposely been chosen from the narrower use of tests with small groups and the use which relates to the practical handling of the individual pupil, rather than to the larger application to administration or research. The teacher cannot well use the refined and elaborate methods which are appropriate for such wider use, and the research student, or the administrator of research departments, does not need the rather elementary treatment which is here given. For these reasons the present chapter has been directed particularly to the needs of the teacher rather than of the more highly trained expert.

GENERAL REFERENCES

The following three books are very useful for the directions they give concerning methods of tabulation and graphic representation, as well as for their treatment of statistical procedure.

1. Otis, Arthur S. *Statistical Method in Educational Measurement.* Yonkers-on-Hudson, New York: World Book Co., 1925.
2. Rugg, Harold. *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Co., 1925.
3. Williams, J. Harold. *Graphic Methods in Education.* Boston: Houghton Mifflin Co., 1924.

Chapter XIII

BASIC FACTS UNDERLYING THE EDUCATIONAL USES OF TESTS

In this chapter the basic facts concerning the prediction of mental growth and ability, individual differences, correlation, and the relation of the characteristics revealed by mental tests to conduct are presented. In the following chapter the educational uses which can be made of tests in the light of these facts will be discussed.

1. The constancy of intelligence and the prediction of intellectual ability

The usefulness of measures of intellectual ability depends in part upon their stability and the possibility of predicting the individual's future intelligence from the measure of intelligence made at a given time. If the purpose of the test, for example, is to classify pupils so that the demands made upon them may be adjusted to their abilities it is necessary that their abilities shall remain more or less constant. If pupils are to be advised to take one course or another according to their ability we must be able to forecast the general level of these abilities. It is important, therefore, to know how constant the measures of intellectual ability actually are.

We may distinguish between the constancy of the I.Q. and the constancy of ability. The I.Q. is a particular measure of ability. In discussing "Scores and Norms" the question has already been raised whether the I.Q. is a comparable measure of ability at successive years, particularly for children at the extremes of ability. It has been shown that the I.Q. is not a perfectly constant measure. This fact

is due to defects in the I.Q. as a relative score. The difficulty would not arise if we used a measure which was not subject to the same criticism as applies to the I.Q., as for example, the standard score. The criticism of the I.Q. does not affect the correlation between scores at successive years but it would affect the prediction of the actual score of a particular child. In the following discussion we shall disregard the question of the suitability of a particular measure of intelligence and consider only the question whether the intelligence of children actually remains constant or fluctuates widely.

In determining whether the child's intelligence fluctuates we must make allowance for the error of measurement in all tests. This factor in variation can be estimated satisfactorily for practical purposes by comparing tests made at short intervals of time. This is done in estimating the reliability of tests. Unreliability will, of course, lower the accuracy of prediction but it will do so not on the ground of the existence of a change in ability but rather because of our inability to measure accurately.

As we have already seen, the reliability of tests varies widely. The correlation of some of the more extensive intelligence tests upon repetition is in the nineties and sometimes is 95 or above. The correlation of many nonlanguage tests and of performance tests is usually somewhat lower. The correlation of tests of pre-school children and of infants is usually still lower. It may be in the sixties or in the seventies.

We may express the degree of unreliability or the degree of variation upon repetition in terms of the average change in score. This procedure has been used in many studies in the case of the I.Q. The average change on immediate repetition is about 5, meaning, of course, that approximately half the changes are more than 5 and half are less

TABLE XXII. MEASURES OF THE VARIATIONS IN THE I.Q. ON RETESTING AS FOUND IN SEVERAL TYPICAL STUDIES

Author	Number of Cases	Percentage Differing 10 Points or More	Limits of Middle 50 Per Cent	Average Change	Coefficient of Correlation Between Two Tests
Terman [1]...	435	.15	− 3.3 to + 5.7	4.5	.93
Rugg and Colloton [2]	137	.12	− 2.3 to + 5.6	4.7	.84
Garrison [3]	468	.085	− 2 to + 4 − 3 to + 4 − 3 to + 5	5.4	.88
Rugg, L. S.[4]	114		− 1.2 to + 1.9	3.1	.95

[1] Lewis M. Terman, *The Intelligence of School Children*, Chapter IX.

[2] Harold Rugg and Cecile Colloton, "Constancy of the Stanford-Binet I.Q. as Shown by Retests," *Journal of Educational Psychology*, XII (September, 1921), 315–22.

[3] S. C. Garrison, "Additional Retests by Means of the Stanford Revision of the Binet-Simon Tests," *Journal of Educational Psychology*, XIII (May, 1922), 307–12.

[4] L. S. Rugg, "Retests and the Constancy of the I.Q.," *Journal of Educational Psychology*, XVI (May, 1925), 341–43.

than 5. The facts concerning the change in the I.Q. on repetition of the test are sufficiently well represented in Table XXII. The later studies have not altered the findings of these earlier ones.

A further question now remains regarding the accuracy of prediction over a longer period of time. The general finding is that the correlation is lower and the prediction less accurate the longer the time interval between the tests.

Thorndike,[1] for example, has summarized the results of thirty-six correlations between first and second tests classifying them according to the time interval between the tests. He derives an equation for the relation between the correlation and the time interval, according to which the cor-

[1] Robert L. Thorndike, "The Effect of the Interval between Test and Retest on the Constancy of the I.Q.," *Journal of Educational Psychology*, XXIV (October, 1933), 543–49.

relation for zero time interval is .889; for thirty months, .814; and for sixty months, .698. A similar drop in the correlation between successive tests was found by Brown,[1] although the drop is not so pronounced. He finds that the average correlation for an interval of one year is .86 and for nine years, .78. The average change in points for one year is 5.36 and for nine years, 9.34. Similar results in regard to the correlation of successive tests were found by Lincoln[2] and Slocombe.[3]

There seems to be some evidence that the I.Q.s of gifted children fluctuate even more widely over longer intervals of time than those of children of average or inferior ability. Nemzek,[4] for example, gives the correlation for five groups of gifted children after one or two years' interval. These correlations range from .53 to .73. Terman,[5] however, disagrees with this finding. He reports correlations ranging from .60 to .81 depending upon the method used in calculating the correlation and asserts that this is about the same relation as found in other studies over an equal interval of time, six years.

Although there is a considerable variation in the standing of individuals within the group after a longer interval of time, say one of five or six years, children do not usually shift from one extreme to the other. Children of the highly

[1] Ralph R. Brown, "The Time Interval between Test and Re-Test in Its Relation to the Constancy of the Intelligence Quotient," *Journal of Educational Psychology*, XXIV (February, 1933), 81–96.

[2] Edward A. Lincoln, "Stanford-Binet I.Q. Changes in the Harvard Growth Study," *Journal of Applied Psychology*, XX (1936), 236–42.

[3] C. S. Slocombe, "Why the I.Q. Is Not, and Cannot Be Constant," *Journal of Educational Psychology*, XVIII (September, 1927), 421–23.

[4] Claude L. Nemzek, "The Constancy of the I.Q.s of Gifted Children," *Journal of Educational Psychology*, XXIII (November, 1932), 607–10.

[5] Barbara Stoddard Burks, Dortha Williams Jensen, and Lewis M. Terman, *The Promise of Youth*. Genetic Studies of Genius, Vol. III. Stanford University, California: Stanford University Press, 1930.

gifted group remain superior, and those who are very dull remain inferior. This is shown by a study of gifted children by Lorge and Hollingworth.[1] These investigators followed up twenty-one children who had been tested at seven to nine years of age and were found to have I.Q.s of 140 or thereabouts. In college their records indicated that they were equal to about the seventy-fifth percentile. They were therefore doing superior work in a highly selected group.

There is some evidence that even larger shifts may occur in the period of infancy or pre-school life. Wellman,[2] for example, has shown that children in the pre-school of the University of Iowa made large gains in I.Q. and that these gains persisted even to the college period. Skeels [3] has shown that foster children whose parentage would lead one to expect an average I.Q. of below 100 have, on the contrary, an average I.Q. of approximately 115. These studies, together with previous studies of foster children and more recent studies of twins, which will be referred to in the chapter on "The Interpretation of Mental Tests," seem to indicate that a larger change in the I.Q. may be produced by influences in the earlier years than most psychologists have hitherto thought possible. This question is one which is still in the realm of controversy, but the present evidence is sufficient to indicate that prediction of the

[1] Irving Lorge and Leta S. Hollingworth, "Adult Status of Highly Intelligent Children," *Pedagogical Seminary and Journal of Genetic Psychology*, XLIX (September, 1936), 215–26.

[2] Beth L. Wellman, "Some New Bases for Interpretation of the I.Q.," *Pedagogical Seminary and Journal of Genetic Psychology*, XLI (September, 1932), 116–26; "The Effect of Pre-School Attendance upon the I.Q.," *Journal of Experimental Education*, I (December, 1932), 48–69; "Growth in Intelligence under Differing School Environments," *Journal of Experimental Education*, III (December, 1934), 59–83.

[3] Harold M. Skeels, "Mental Development of Children in Foster Homes," *Pedagogical Seminary and Journal of Genetic Psychology*, XLIX (September, 1936), 91–106.

I.Q. in infancy or early life is much less certain than its prediction in later years.

There are three possible explanations of the greater variation in intellectual ability after a longer interval of time than after a shorter period. The first hypothesis is that the difference is not real but is due to the fact that there is less in common in the test for widely separated years than for adjacent years. This would be true for tests like the Binet scale but not to the same degree for tests in which the same kind of material is used throughout a wide range.

The other two hypotheses rest upon the same general conclusion that the change is real, but give a different explanation of its origin. The explanation advanced by Terman is that the change is due to a difference in the inherent growth rates of different children. The other explanation is that it is due to the effect of environment which is more favorable to some children than to others.

Whatever the explanation may be the fact that accuracy of prediction is limited and becomes less over a longer period of time must be taken into account in the practical use of tests.

To summarize, tests may be used for prediction if we make due allowance for the error of such prediction. Over a short interval of time the error of prediction is due to the unreliability of the test, with an average error of about five points in I.Q. Over a longer interval of time the accuracy of prediction becomes progressively less until the average error amounts to perhaps ten points in I.Q. The prediction of the later intellectual status of infants and pre-school children is much less certain than a prediction of the later status of children of school age. Prediction is progressively less accurate the younger the child is. This inaccuracy in the case of young children may be due to the inaccuracy of the test or to the fact that the test measures different forms

of development from those measured by the later tests. The later tests measure performances which are more nearly like those acquired in the school and in many occupational activities. That is, the tests may be more adequate measures of the abilities which we wish to predict. In addition to this, however, the child's intellectual ability may have actually become more stable. If the environment affects the child's intelligence we should expect early environment to be more effective than later environment. Once the child's intellectual habits become fixed it may be more difficult to change them. These reasons may combine to make later prediction more accurate than early prediction.

2. Individual differences

The fact that extreme individual differences in mental capacity exist has, of course, appeared repeatedly during the course of our discussion. They were revealed, for example, in the distribution of the I.Q.s, in the chapter on the Stanford Revision of the Binet scale, and illustrations were given in the chapter on the tabulation of mental test scores. The magnitude of individual differences is so much a matter of common knowledge that it is hardly necessary to dwell upon its existence. We may give but one illustrative statement. According to Terman's calculation of the percentage of children of various I.Q.s, we may calculate the number of those of twelve years of age whose general intelligence gives them a mental age a specified distance above or below twelve years. If we tested all twelve-year-old children we would find that 10 per cent of them had a mental capacity equal or inferior to that of the average child of ten years and two months. At the other end of the scale the brightest 10 per cent of children would be found to have mental age equal or superior to that of the child of thirteen years and eleven months. In other words, 20 per cent of the chil-

dren would be either approximately two years inferior or two years superior to the average twelve-year-old child. The average of the upper tenth and the lower tenth would be separated from each other by a space of four years. In a school of one thousand children, two hundred would belong to one or the other of these two extreme groups.

Our knowledge of differences in general intelligence is more complete than our knowledge of differences of special intellectual capacities or in the non-intellectual traits, such as emotion, will, or moral character. Our methods of measuring general intelligence are also better developed than are our methods of measuring most of these other traits. Our knowledge is sufficient, however, to indicate clearly that differences in the other traits are of sufficient importance to merit serious consideration. Where we cannot as yet measure them accurately, we should estimate them to the best of our ability.

Various particular intellectual capacities are to a considerable extent specialized. A pupil may have high general intelligence and yet may be poor in ability to do manual work, or may be very deficient in musical capacity. We cannot classify pupils in these subjects merely upon the basis of a general intelligence test. Furthermore, the pupil may be comparatively low in general intelligence and yet may have unusually high capacity in some specialized direction. This high capacity may constitute the pupil's chief educational and vocational opportunity. To overlook it and to fail to give the pupil the appropriate training would be a serious blunder on the part of the school.

Again, traits of character, temperament, or will are important factors in determining the pupil's success in school or in life, and require both recognition and training. A pupil of very low intelligence cannot, by the exercise of any amount of resolution or energy, raise himself above a medi-

ocre level of academic accomplishment, though he may do a much better grade of work than the average of his intelligence group. A person of high intelligence, on the other hand, may fail utterly in achievement because of an unstable emotional life and a poor adjustment to his social environment.

Besides these individual differences in mental traits, other differences frequently affect the pupil's school work. His physical condition may impair his capacity for work. His home environment or his childhood associates may be either favorable or unfavorable to the development of intellectual interests and to consistent achievement. The rate of physical growth probably has some bearing upon the child's mental development, and upon his social attitudes. Just how important this factor is we do not know. It probably has some influence in determining the group with which a child can associate upon equal terms.

3. Correlation between mental tests and other measures of capacity or achievement

We have already seen that intelligence tests correlate fairly closely with one another, and that it is largely because of this that we judge them to measure general mental capacity or intelligence. Two general intelligence tests, if given to a class of fifty to one hundred pupils, usually give scores which correlate with one another from .70 to .80. This indicates that they are measuring something, and that this something is common to the two tests.

The next question is, Does this something which is measured by this test agree with what we ordinarily mean when we use the term brightness or intelligence? Does the test agree with our judgment of intelligence? In general, there is agreement, but we find great variation in the correlation between tests and the judgment of various individuals.

The judgments of some persons have very low correlation, and the judgments of others have comparatively high correlation. The correlation between teachers' judgment and mental tests may run from as low as .30 to as high as .60, or even higher. Some persons have a clear idea about what is meant by general intelligence, and are good judges of it; other persons either have a vague idea, or are poor judges of individuals. On the whole, the tests, being more consistent, are to be relied upon more implicitly than are judgments. The variations which are found between the various cases in which mental tests are correlated with judgment may be explained in large part by the differences in the training or ability of the judges.

When we come to the correlation between mental tests and educational achievement, the measures with which we are dealing are more objective. Several illustrations will serve to bring before us the typical facts. Correlations between individual mental tests and group tests with composites of educational achievement are reported by Gates.[1] It will be seen that mental age, as measured by the Stanford Revision of the Binet scale, correlates about as closely with achievement as do scores in the verbal group test. The non-verbal group tests, however, correlate less closely with achievement. This is perhaps due to the fact that school achievement is based more upon verbal than upon non-verbal type of performance. The very low correlations of the non-verbal test with achievement in the higher grades is probably due to the fact that only one group test was used in these grades, and that one a test which has shown uniformly low correlation with school achievement. In general, we may say that the correlation of intelligence tests

[1] Arthur I. Gates, "The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables," *Journal of Educational Psychology*, XIII (March, April, and May, 1922), 129–39, 223–35, 277–85.

TABLE XXIII. THE CORRELATION OF INTELLIGENCE TESTS AND SCHOOL ACHIEVEMENT

GRADE	1	2	3
	Achievement with Mental Age (Stanford)	Achievement with Verbal Group Tests	Achievement with Non-verbal Group Tests
I	0.36	....	0.30
II	0.44	....	0.23
III	0.47	0.65	0.22
IV	0.42	0.54	0.22
V	0.51	0.49	0.17
VI	0.67	0.57	0.29
VII	....	0.52	0.08
VIII	....	0.47	−0.15

with composite school achievement in the elementary school, as shown in this investigation, is in the neighborhood of .50. An illustration of the results of the application of tests in the high school may be found in a report by Proctor.[1] Proctor found the correlation between the I.Q. and the composite of school marks to be .545.

Something of the variation in the correlation found in different institutions by different investigators may be gathered from typical statistics from the college field. At Yale, Anderson applied the Army Alpha test to four hundred freshmen, and found the correlation between composite standing and the test to be .377. The correlations in Table XXIV are reported by a committee of the faculty of Stanford University under the chairmanship of Terman.[2]

Jordan, at the University of Arkansas, who gave the Army Alpha to 315 college students, reports a correlation

[1] W. M. Proctor, "The Use of Intelligence Tests in Educational Guidance," *School and Society*, VIII (Oct. 19, 26, 1918), 473–78, 502–9.

[2] L. M. Terman and Others, *Report of Sub-Committee of Committee on Scholarship on Student Ability*. Stanford University Press, 1923.

TABLE XXIV. THE CORRELATION BETWEEN ACHIEVEMENT IN COLLEGE AND INTELLIGENCE TESTS

Stanford University	
275 freshmen men, scholarship for three quarters	.54
58 freshmen women, scholarship for three quarters	.63
677 freshmen men, scholarship for three quarters	.49
204 freshmen men, scholarship for six quarters	.48
30 freshmen women, scholarship for six quarters	.67
138 transfer men, scholarship for six quarters	.42
35 transfer women, scholarship for six quarters	.49
Columbia, 199 New Plan men, scholarship for 1 semester	.60
Columbia, 111 New Plan men, scholarship for 2 years	.67
Columbia, 122 Old Plan men, scholarship for 2 years	.50
Mills, 157 women, scholarship for 1 year	.70
Brown, 300 men, scholarship for 1 year (?)	.60
University of California, 273 men and women, scholarship for 1 year (?)	.47
Goucher College, 243 women, scholarship for 1 year (?)	.60
Trenton Normal School, women, scholarship for 1 semester	.56
University of Pittsburgh, 569, both sexes, scholarship for 1 semester	.51

with college standing of .485. Wood reports a correlation between the Thorndike College Entrance Intelligence Examination and the two-year scholarship score of .594. He reports that the Thorndike examination correlates with points earned by 106 students .454. Colvin reports that in the case of three hundred students in Brown University the correlation between the Colvin test and college grades was .60.

These correlations are typical. The correlation between intelligence tests and composite standing of the pupils may be said, then, to lie usually between .40 and .60. Probably, in the majority of cases the correlation will be found to be in the neighborhood of .50, but under very favorable conditions it may be somewhat above this.

The practical meaning of this correlation is that it enables us with a moderate degree of accuracy to predict the grade of work which a student will do in school or college. Two questions confront us in an attempt to evaluate and apply

this fact. In the first place, how does the accuracy of prediction or the closeness of correlation of intelligence tests compare with the predictive value or the correlation of previous school work? Consider first the correlation between average standing in high school and in college. The correlations which have been reported vary considerably. Wood reports, in three cases, a correlation between secondary school marks and college scores of .262, .331, and .15.[1] Thorndike,[2] in an early study, reports correlation between college entrance examinations and marks in the four college years of .62, .50, .47 and .25 respectively. Dearborn, in his Wisconsin study, reports a correlation of .80.[3] The very low correlations reported by Wood are probably due to the variation in the marking standards of different institutions. They would be very much raised if allowance were made for these variations, or if a common standard were used. The very high correlation by Dearborn is difficult to explain. It probably is not typical, however. We may assume about .50 as a typical correlation between high-school standing and college standing under favorable circumstances. This means that standing in high school has about the same predictive value for college standing as have intelligence tests.

The second question to be raised is, How are these correlations to be interpreted? The pupil's standing in the intelligence test and in school or college work may differ for two causes. In the first place, the two may depend upon different capacities. In the second place, the inaccuracy of the two measures may reduce the correlation between

[1] Ben D. Wood, *Measurement in Higher Education*, pp. 85–86. Yonkers-on-Hudson, New York: World Book Co., 1923.

[2] Edward L. Thorndike, "An Empirical Study of College Entrance Examinations," *Science*, XXIII (June 1, 1906), 839–45.

[3] W. F. Dearborn, *Relative Standing of Pupils in the High School and in the University*, p. 21. University of Wisconsin Bulletin No. 312. 1909.

them. We have already seen that the correlation between two intelligence tests, which presumably measure about the same ability, is between .70 and .80, rarely going beyond the second figure. This represents roughly the correlation we get when inaccuracy is the chief disturbing factor. A comparable figure for school marks may be found in the correlation between the grades of students in the freshman year and the sophomore year in college. This correlation is reported by Wood to be .72.[1] From this it appears that the composite of a year's marks, at least at the college level, and presumably at the high-school level, is almost as accurate as intelligence test scores. If intelligence tests and marks measure exactly the same thing, then, we should expect them to correlate with one another about as closely as the marks of one year correlate with the marks of another, or as one intelligence test correlates with another, namely, between .70 and .80. When correlations are lower than this, we may conclude that the marks and the tests measure somewhat different capacities. Marks, for example, depend not only on intelligence, but also upon previous training, industry, and interest. The lower correlation between the high-school and college marks than between marks of two college years may be ascribed either to the fact that high-school work and college work, being of a somewhat different character, demand somewhat different abilities, or to the fact that the marking standards of different institutions vary so largely that there is larger error in comparing them than in comparing the marks of different courses in the same institution. Entrance examinations, it may be noted in passing, have about the same correlation with college grades as does the intelligence examination.

In order that we may have before us in somewhat more concrete form than is represented by the correlation co-

[1] Ben D. Wood, *op. cit.*, p. 133.

efficients the relationships between the intelligence test score, the marks on entrance examinations, and the average mark of previous school work on the one hand, with standing in college on the other, three correlation tables are presented. The illustrations are selected from high-school and college work, because our data for this level are more complete than for the earlier grades. The relationship is similar, however, to the relationship between standing in the elementary school and in the high school.[1] The entries of these correlation tables represent percentages rather than numbers of cases. They are so arranged that each column and each row add up to 100 per cent. The tables are to be interpreted thus: In the case of Table XXV A, 49 per cent of the pupils who are in the lowest quarter of the class in their standing in the intelligence test, represented by the first column to the left, are also in the lowest quarter in their standing in college, represented by the horizontal row at the bottom. By running up the first column on the left, we find that 38 per cent who are in the lowest quarter of the intelligence test are in the second quarter in their college standing, 10 per cent are in the third quarter, and 2 per cent are in the top quarter. The percentages in the squares along the diagonal from the lower left to the upper right-hand corner represent students who stand in the same quarter according to the tests and to their college marks.

Suppose, now, we were to use the intelligence test as a means of prediction and of classifying the students into sections. If four sections were formed, from 31 per cent to 57 per cent of each group would be properly placed as judged by their marks, from 18 per cent to 38 per cent of each group would be in a section one removed from their proper place, from 6 per cent to 16 per cent would be in sections two re-

[1] Cf. on this point John Addison Clement, *Standardization of the Schools of Kansas.* Chicago: University of Chicago Press, 1912.

TABLE XXV A. CORRELATION BETWEEN INTELLIGENCE TEST SCORE AND STANDING IN COLLEGE

Standing in College \ Intelligence test				High
	.02	.16	.25	.57
	.10	.18	.44	.29
	.38	.31	.25	.06
Low	.49	.35	.06	.08

TABLE XXV B. CORRELATION BETWEEN ENTRANCE EXAMINATION MARK AND STANDING IN COLLEGE

Mark in Freshman Year \ Entrance mark				High
	.14	.09	.23	.54
	.14	.16	.49	.20
	.26	.35	.18	.20
Low	.47	.40	.09	.05

TABLE XXV C. CORRELATION BETWEEN GENERAL STANDING IN HIGH SCHOOL AND STANDING IN COLLEGE

Standing in College \ Standing in High School				High
	.042	.135	.178	.644
	.221	.212	.398	.169
	.280	.314	.263	.144
Low	.458	.334	.161	.042

moved, and from 2 per cent to 8 per cent would be in sections three removed from the correct one. The correlation of .50, which was found in this instance, represents the degree of accuracy in prediction and in classification shown in this table, when the classification is made into four groups. If classification were made into three groups, the accuracy would be somewhat higher. This gives us an idea of the practical value of the intelligence test and of other means of predicting the pupil's standing and of classifying him.

The criterion which is here used is accuracy of prediction, but accuracy of prediction is not the only criterion to use. The intelligence test, by measuring somewhat different capacities from those which are measured by school marks, may give us a partial basis for analyzing and explaining a pupil's achievement or failure.

If the mental test measures one or more factors in educational achievement, and if the previous attainment of the student represents another factor or group of factors, we might expect a combination of these two measures to give a more accurate prediction than either one alone. There seems to be some evidence that this is the case. A number of studies have been made, for example, in which the several factors are represented in a regression equation, with the prediction made on the basis of this equation. An example of these studies is the one made by Blair.[1] The accuracy of prediction of the intelligence test alone is represented by correlation of .491 with the average of first-year grades in college, with previous attainment by correlation of .631 (numerical grades) and .599 (letter grades), and a combination of the two by correlation of .667 (numerical grades) and .641 (letter grades). This does not mean that regression

[1] John Lewis Blair, "Significant Factors in the Prediction of the Success of College Freshmen." Unpublished Doctor's thesis, University of Chicago, 1931.

equations should be used under ordinary school conditions in making predictions. It does mean, however, that both factors should be taken into account.

4. The value of mental tests as measures of particular factors in achievement

There is evidence that intelligence tests measure certain components of the abilities required in school work more than they do other components. This has been shown by analyses of the causes of failure of students. In many cases it is found that the failure is not caused by lack of intellectual ability but by other deficiencies. In such cases the intelligence test enables us to determine whether or not the failure is due to intelligence deficiency or whether it is necessary to look to some other cause. Statistical evidence that the intelligence score is a measure of only one of the various components required in school achievement is found in a study by Pressey, which may be taken as an example.[1] Pressey studied 116 junior-high-school students with the purpose of finding out the factors in their school success. His method was to have the students rated by the teachers on health, school attitude and preparation, and intellectual ability, and then to correlate these various ratings with marks. A significant finding was the partial correlation between marks and ability, on the one hand, and between marks and school attitude, on the other hand. The partial correlation of ability with marks was .49, and of school attitude and marks was .43. This means that if the pupils were all equal in school attitude marks would correlate with ability to the extent of .49. If they were all equal in ability,

[1] S. L. Pressey, "An Attempt to Measure the Comparative Importance of General Intelligence and Certain Character Traits in Contributing to Success in School," *Elementary School Journal*, XXI (November, 1920), 220–29.

marks would correlate with school attitude to the extent of .43. In other words, ability makes a contribution to school achievement independent of attitude, and attitude makes a contribution which is independent of ability. It is desirable to have a separate measure of each of them, in order that we may analyze a pupil's performance and determine what contributes to his achievement or success. Intelligence tests are important, then, because they help make this analysis possible.

If factors of personality in addition to ability are important determinants of the pupil's attainment in school we might suppose that personality tests could be used to supplement the tests of ability as a basis for predicting the child's attainment. Such, however, is not yet the case, whatever may be the future application of tests of personality. These tests, at the present time, help us to analyze the attitudes and emotional stability of individual children and, in many cases, to discover facts which are useful in assisting them to attain greater emotional poise and better adjustment to their social environment. In other words, the tests are of diagnostic and therapeutic value but are not of value as a basis for the routine administration of the pupils. Whether they ever can be used in this way may be a question. At any rate, they are not suitable for such use at the present time.

5. The relation of intelligence or ability to conduct

Soon after the introduction of the Binet scale in the United States, intelligence tests came to be used as part of the basis for the interpretation of the conduct of delinquents. A person convicted of committing an offense would be given an intelligence test in order to determine the degree of his responsibility. The intelligence test score was commonly given in terms of mental age and the indi-

vidual's responsibility was estimated to correspond to that of the average child of corresponding mental age. Intelligence was thus assumed to have a close relation to conduct and to responsibility for conduct.

The principle involved in this interpretation of court cases was extended by many to cover the whole range of mental ability in its relation to conduct. For example, intelligence was conceived of as the ability to forecast the consequences of behavior and this ability to forecast the consequences was regarded as the chief basis for the control of behavior. Hence, intelligence was inferred to be the largest factor in behavior and to be a reliable basis for prediction of behavior.

This belief in the close relation between intelligence and behavior was apparently confirmed by statistical studies which indicated, first, that delinquents have a lower average intelligence than children who are not delinquent or adults who are not delinquent, and, second, that there is a larger percentage of delinquents among those of low intelligence than among those of normal or superior intelligence. These conclusions were based upon the estimate of intelligence by means of the Binet scale.

This conclusion was brought into question by two lines of investigation. The first consisted in a comparison of the intelligence of delinquents with children of the same social or economic class. It is known that delinquents, at least those who are committed to institutions, come more largely from poor districts, such as slum areas, than from good districts. It has also been established that the average intelligence of the inhabitants of poor districts is much lower than that of those of good districts. If, now, we compare the intelligence of delinquents with that of children from the poor districts we find that it is not inferior to the average.

The second line of study was to compare the intelligence

of adult prisoners with the intelligence of unselected groups of men as represented by the recruits of the Army. When this was done by means of the Army Alpha, it was found that the average intelligence of prisoners was not lower than that of the population as a whole. It was found, however, that the kind of crime committed was related to the intelligence of the person committing the crime. For example, crimes of violence are more commonly committed by persons of low intelligence, whereas crimes against property, such as embezzlement and forgery, are more often committed by those of high intellectual ability. These investigations appear to indicate that the relation between intellectual ability and conduct is not of the simple and direct sort it was once thought to be.

The relation between intelligence and conduct is an involved one. Conduct does, to be sure, depend in part upon the foresight of its consequences. A clear knowledge of the consequences of one's acts, however, is difficult to obtain. Many errors may be made in forecasting the consequences and large difference of opinion may exist regarding it. Furthermore, the consequences are not invariable but are subject to a variety of factors. It is common, therefore, for an individual to believe that he can escape the consequences which commonly follow a given set of actions. That is, he believes that he may be an exception to the rule.

It is necessary to recognize also that conduct is determined by desires, primarily, and that the estimation of consequences is largely in terms of the expectation of the satisfaction of desires. Our judgments are notoriously influenced by our desires, an observation true of the intelligent as well as of the stupid person. The strength of desires and the extent to which they are satisfied by legitimate means are therefore important factors in conduct.

Again, an undue emphasis upon intelligence as a factor in conduct leaves out of account the influence of the social group and of the customs, mores, or standards of the group upon the conduct of the individual. These standards are impressed upon the individual from his early years and become incorporated into habits which constitute a strong controlling force in conduct. The factors in conduct, then, are complex, and intelligence is only one of them. It is probably a more important factor in the case of feeble-minded individuals than it is in persons of normal or superior intelligence.

The relation of intelligence to conduct is pertinent in the interpretation and control of the child's behavior in school as well as of the individual's behavior in society at large. The relation of intelligence to school behavior is probably more an indirect than a direct one. It is, of course, a large and perhaps the chief factor in the child's general school achievement. If the child's intelligence is not sufficiently taken into account in studying the tasks he is asked to perform he may resort to misbehavior to relieve the emotional conflict which ensues. The misconduct which thus arises out of maladjustment between the child's ability and his tasks may be removed by setting tasks which are in conformity to his ability. Intelligence tests, therefore, are not by themselves either the means of predicting conduct or of interpreting it. Taken in conjunction with other facts, however, they may be useful for these purposes.

Nothing that has been said should be construed to mean that character education has no intellectual content or basis. It implies only that individual differences in intellectual ability are not the main factor in differences in conduct. It does not imply that the conduct of an individual does not depend in large measure upon his conceptions of what good conduct should be or upon his under-

standing of the relation between himself and other persons. While the possession of such understanding does not guarantee that a person will act in accordance with it, since conduct is dependent upon emotional factors and other factors in the situation, nevertheless a person cannot well act appropriately in a given situation unless he has an adequate conception of its meaning. The improvement of conduct, then, depends upon the improvement of understanding, and the gaining of correct understanding is the first step in the development of conduct which is adequate to the situation.

The question may be raised as to the relation of other characteristics than the intellectual to conduct and the relation of scores on personality tests to behavior. As in the case of the relation of intelligence to conduct, a common early misconception concerning the relation of constitutional personality to conduct may be ruled out. This is the notion that there exist certain persons who lack a fundamental moral sense and who are therefore incapable of moral judgment and of conduct in conformity to it. It has been believed by some criminologists and psychologists that certain persons are moral imbeciles, that is, lack capacity to make moral distinctions. This opinion has been generally abandoned. It is an oversimplified way of explaining the differences in conduct. There may be a relation between natural temperament and conduct but if there is it is an indirect and complex one as in the case of the relation of intelligence to conduct.

Differences in temperament probably exist, but they affect conduct only because they affect the way one reacts toward particular persons or in particular situations. They may thus be the occasion of conflict between persons which creates problems of conduct. They do not, however, predetermine that a person shall act in a particular way in a

given situation. Furthermore, they do not bring about conduct of a certain general sort but only influence conduct in particular cases. We cannot then say, in general, that a given temperament predisposes to good conduct and another to bad conduct. The result is entirely a matter of adjustment to particular situations. It is true that a given kind of situation involving a problem between a given person and another may arise frequently enough to lead to an undesirable habit of conduct in the individual concerned. The resulting habit, however, is a function of the recurring situation as much as of the constitutional nature of the individual.

Because there is no such thing as general moral or immoral constitution there is no such thing as a test of general moral or immoral constitution. Some tests have a bearing upon the interpretation of the conduct of individuals and may be used helpfully to make such interpretation. There exist, in the first place, as indicated in the chapter on "Tests of Personality Traits," tests of specific conduct forms, as tests of honesty in particular situations. These may be used to make a record of the individual's behavior but they should not be interpreted as measures of general traits. Again, there are tests of neurotic disposition, and of other characteristics of personality which may reveal facts about the person useful in interpreting his conduct in given situations. The tests are useful in giving information about the individual which may be used in the interpretation of his conduct, but they always require interpretation in the light of the entire situation. They cannot be taken simply as measures of moral or immoral disposition. No such tests exist.

The interpretation of tests with respect to conduct, therefore, is a somewhat less direct and more complex matter than is the interpretation of tests of ability. For this

reason they are not subject to the same routine use as are ability tests. For the present, at least, they will be used mainly by clinical workers who have been trained to analyze the behavior of individuals and to interpret it in the light of their life history, home background, ability, and personality.

Chapter XIV

THE EDUCATIONAL USES OF TESTS

1. The general intelligence level and its relation to achievement

In Chapter XIII were considered the basic facts which should guide the application and use of mental tests in the school. In this chapter will be taken up the various aspects of the school program for which mental tests can be helpfully used.

Mental tests are helpful in dealing with students both as individuals and as members of a group. In dealing with the individual pupils tests are useful in answering such questions as, when is the right time for the child to enter school; how may a pupil be kept in best adjustment to his work; and how may students best be selected for college or professional school. In dealing with the pupils as a group, the administrator is faced with the need of a sound basis for classifying them into ability groups, in selecting those who should be put into special classes, and in setting up sound standards for a guidance program. The problem of homogeneous grouping hinges on the answer to the question of the relation of the I.Q. and school achievement. It is with these questions that this chapter will deal.

While not very much use has been made of the fact, we have clear evidence that there are noticeable differences between communities in the average standing of their children in intelligence tests. We may cite merely one example reported by Pintner.[1] Pintner gives the comparative rating of the children in a town in Ohio and one in Kansas. The median mental index of the children in the Ohio town is 40, which is ten below normal, while the median index of the Kansas children is 51, which is one above normal. Similar

[1] Rudolf Pintner, *Intelligence Testing*, p. 239.

marked differences have been found between rural children as a group and city children as a group, and also between children in one section of a city and in another section of the same city. Such differences as these may be due partly to native or inborn differences in capacity, and partly to differences in early training, but in any case they represent differences in present capacity to do school work. They are therefore significant because they constitute one basis for the interpretation of the achievement of the children.

The use of the average intelligence of the children of a community to interpret the results of achievement tests may be illustrated by an example. In a certain state the children of a group of cities were given an intelligence test and also the Woody Arithmetic Test. We have the scores of the children in comparison with the norms in both the intelligence and the arithmetic tests. The facts are given in Table XXVI.

TABLE XXVI. THE SCORES OF CHILDREN IN A GROUP OF CITIES IN AN INTELLIGENCE TEST, COMPARED WITH THE NORMS

School Grade	III	IV	V	VI	VII
Scores	38	58	77	92	103
Norm	40	60	78	96	110
Difference	−2	−2	−1	−4	−7

THE SCORES OF THE SAME CHILDREN IN THE WOODY ARITHMETIC TEST, COMPARED WITH THE NORMS

School Grade	III	IV	V	VI	VII
Scores	9.4	12.4	14.4	15.2	15.9
Norms	9.7	12.7	15.5	17.8	18.5
Difference	−.3	−.3	−1.1	−2.6	−2.6

It is apparent from an inspection of the table that the children in these cities make scores upon the intelligence tests only slightly lower than the norm, though in the sixth and seventh grades the inferiority is slightly greater than in the third to the fifth grade. In the arithmetic test, the scores are practically equal to the norm in the third and fourth grades, but become inferior in the fifth grade and markedly inferior in the sixth and seventh grades. The proportional inferiority in the arithmetic test in the upper two grades is much greater than the inferiority in the intelligence tests. We may therefore conclude that the teaching in these upper two grades is less efficient than in the lower grades, or that there is less emphasis given to the subject, or that some other circumstance operates to lower the children's achievement below what we should expect it to be. Comparisons of this sort may be used to locate the spot in the school system which needs special supervisory attention.

It has become commonplace in the reports of school surveys to point out variations among schools, and among classes within the schools, in the achievement of children. Similar variations may also be found in the average intelligence rating of schools or of classes. It is not necessary to give illustrations, since the principle is similar to the one brought out in the preceding paragraph. It is quite evident that when variations in the achievements of children of a school or class are found, it will be very helpful in interpreting the causes of such variation to know the intelligence rating of the specified group. While there is danger in attempting to determine with too much exactness what the achievement of a group of children should be from their intelligence scores, nevertheless, gross differences can readily be interpreted by help of them.

It is possible by means of intelligence tests to secure facts which are of assistance to the principal, the supervisor, or

the superintendent in judging the work of individual teachers. Besides using the test scores to interpret the achievement of the pupils under a teacher's care, they may be used to estimate the accuracy of the teacher's judgment of the abilities of pupils and to gain light upon the basis of the teacher's marks. The teacher's success in handling pupils will depend to a considerable extent upon how accurately she judges their ability. We have seen that different teachers vary considerably in the accuracy of their judgment. To overestimate the capacity of a pupil will result in applying undue pressure to him; to underestimate the ability of the pupil, on the other hand, may result in the failure to stimulate him to as good work as he can do.

2. Administrative use of mental tests in dealing with individual pupils

Enough has been said to indicate that the score on a mental test is rarely if ever to be taken as the sole basis for a decision regarding the pupil. Responsible psychologists and educators usually emphasize the fact that mental tests are only one means of judging the pupil. Dickson gives the following list of items as necessary in order to deal with a pupil intelligently:[1] (1) chronological age, (2) mental age, (3) intelligence quotient, (4) grade, (5) accomplishment in school work, (6) application or industry, (7) health, (8) home environment, (9) nationality and language difficulty, (10) special or unusual conditions bearing upon school success. The treatment of the individual pupil is always a complex problem. Mental tests furnish valuable aid to the solution of this problem, but they must always be interpreted in the light of all the facts which can be gathered about the pupil.

Keeping this principle in mind, we may now list the vari-

[1] Virgil E. Dickson, *Mental Tests and the Classroom Teacher*, p. 99. Yonkers-on-Hudson, New York: World Book Co., 1926.

ous ways in which mental tests may be used in the administration of the individual pupil.

3. Mental tests as an aid in the determination of the right time to enter school

It is a well-established fact that when pupils enter school, at the age of six, they are very differently equipped to do successfully the work of the first grade. Out of 76 children who were tested in the kindergarten by Dickson, 24 or 31.6 per cent failed to make normal progress during the subsequent two years.[1] Of 95 children who were tested in the low first grade, 45, or 47.4 per cent, failed of normal progress during the next two years. Of 90 children in the second half of the first year who were similarly tested, 60, or 66.7 per cent, failed in normal progress. Of these 261 children, however, only three of those who had an I.Q. above 110 in the test failed to make normal progress. Of the entire 129 who failed of promotion at least once, 84 had an I.Q. below 90, and only 32 had an I.Q. between 90 and 109. When it is remembered that only 20 per cent of children in general have an I.Q. below 90, the preponderance of the retarded children in this group is very significant.

Superintendent Saam, of Council Bluffs, made an experiment in which children were promoted from kindergarten to the first grade on the basis of their intelligence quotients.[2] The results of the experiment are reported in the following words:

> In an attempt to check up young children who are promoted into the first grade upon the basis of their high quotient, an oral

[1] Virgil E. Dickson, *The Use of Mental Tests in School Administration.* Board of Education Monographs, No. 4. Berkeley, California: Board of Education, 1922.

[2] Theodore Saam, "Intelligence Testing as an Aid to Supervision," *Elementary School Journal*, XX (September, 1919), 26–32.

reading test similar to the Gray Oral Reading Test was given by the primary supervisor in January, 1919, to every child who had entered the first grade in September, 1918. There were 408 students tested. Of the 408, 128, or 31 per cent were rated as superior readers in this test. Of these 408, 35 had been promoted to the first grade at five years of age, because they had a quotient of 115, or over. Of the 35 students with a quotient of 115, or over, 22 or 63 per cent were rated as superior. If conclusions could be drawn from this one test, it would be safe to assume that children five years old with quotients of 115 or over would do the first-grade work better than the unselected six or seven-year-old children.

It is evident that the bright younger children are capable of doing the work of the first grade even better than the average six-year-old child. It is further evident that the dull older children are incapable of doing successfully the work of the first grade or two as the grades are now constituted. Should the bright child be accelerated by being put ahead of those his own age in school, and the dull child be retarded, or should all the children of the same age be allowed to enter school together, and then the work be differentiated according to their capacity?

This brings us to the question whether it is better to advance children of different capacities through the school or through the curriculum at different rates of speed, or whether it is better to attempt to enrich the curriculum for the bright children, and give a simplified curriculum to the dull children, but carry them through it at the same rate. This is not the place for an exhaustive discussion of this administrative problem. We shall recur to the problem and attempt briefly to sum up the considerations for the two types of treatment.

It is evident from this and many other studies that children of school age vary greatly in their ability to do the work of the first grade. It is also clear that tests given upon entrance to school enable us with fair accuracy to

predict which children will do the work of the first grade according to ordinary expectancy and which ones will not. A test which has been found useful for this purpose is the Metropolitan Readiness Test listed in the chapter, "Survey of Point Scales."

After the abilities of children have been determined, however, it is not entirely certain what adjustment should be made to the differences which are found. A method which has been suggested is to allow brighter children to enter the work of the first grade at an earlier age than usual and to hold back the duller children until a later age. This method involves keeping the requirements at a constant pitch and varying the time at which children encounter these requirements. Another method is to allow all children to begin at the same age but to vary the requirements for those of different abilities. A disadvantage of the first method is that it gives the children who have less learning ability a shorter time in which to do the work of the school than is afforded the brighter children. This consideration would seem to dictate the practice of permitting the slower children to enter school at least as early as the brighter children and to adapt the difficulty of the work to their ability.

4. Classification into ability groups

One method of treating children of different degrees of ability at the beginning of their school career, as has already been said, would be to allow them to enter when they have reached a given mental age. The assumption underlying this procedure is that children who are equal in mental age are able to do the same character and quality of school work. This procedure would be the first step in the classification of pupils according to mental age. This classification might conceivably begin at the first grade and be carried forward

throughout all the grades, and the proposal has been made that this be done.

There are several difficulties with this classification by mental age, which we may call vertical classification, since it involves placing the children at a given point in the scale of mental development.

There is first the practical difficulty that pupils do not, as a matter of fact, enter school at the same mental age. They enter at the same chronological age of six, and it is not likely that any wholesale modification of this practice will be adopted in the near future. Some pupils, then, start with a handicap, others with an advantage, and it is not possible to bridge the gap between them.

This one fact alone seems to make necessary a horizontal classification of pupils according to their intelligence. This horizontal classification is based upon the intelligence quotient or some other measure of relative brightness, rather than upon mental age. It means dividing pupils in the first grade, or in succeeding grades, into groups. This procedure is illustrated by organizing three groups of pupils, one containing the bright pupils, another the average pupils, and a third the slow pupils.

The second and more fundamental reason why classification on the basis of the mental age may not be satisfactory is that, even if we should start pupils together in the first grade who have the same mental age, they would not remain equal in mental age. We saw in the chapter on mental development that, if we accept the ratings which are obtained with the Binet scale, children of the same mental age exhibit a wider spread in chronological age as they grow older. Children of high I.Q. gain more rapidly in mental age than average children, and children of low I.Q. gain less rapidly. Group tests indicate much less divergence in curves of mental age than the Binet scale, but they indicate that there is some divergence.

The third objection is that a merely vertical classification places together in the same group children of rather widely divergent chronological ages and stages of physiological and social development. A difference in chronological age of two years at entrance to school would be represented by a still greater difference after the children have been in school six years. It is generally agreed that a wide divergence in ages of children in the same grade is disadvantageous.

A fourth objection has already been referred to, namely, that deferring the entrance of the duller children into school gives them less time to learn, whereas they need more time.

The horizontal classification of children into groups of similar ability may be begun in the first grade, and continued throughout the child's schooling, or it may be begun at some later period. Such grouping has been most commonly carried out in the high school. It has recently been tried out, however, in the lower grades of the elementary school. Probably the most extensive experiment of grouping at this level is the one in Detroit.[1]

Homogeneous grouping may be used to provide opportunity for proceeding at different rates of progress, or to provide enrichment for the bright pupils and a simplified curriculum for the slow ones. Segregation is a general administrative device which provides the opportunity for various sorts of adjustments in curriculum and method.

The differentiation which consists in taking the brighter group through the curriculum more rapidly than the backward group needs no explanation. It represents a type of

[1] Charles S. Berry, "The Classification by Tests of Intelligence of Ten Thousand First-Grade Pupils," *Journal of Educational Research*, VI (October, 1922), 185–203.

acceleration which does not have the disadvantages attendant on skipping grades. It gives the bright child more difficult work than is encountered by children of less ability. It also brings the gifted child to the threshold of high school and college sooner than has been the accepted age.[1]

Many, perhaps most, educators and psychologists consider a qualitative adjustment preferable to this quantitative one. Instead of varying the rate of advancement they consider it better to enrich the course of study for the gifted child. This would seem to imply a corresponding impoverishment of the curriculum for the backward child. It may be that an advantageous qualitative differentiation can be made. For the most part, however, the adjustments which have actually been made consist largely in giving the bright child the sort of work which is part of the regular course of study for a later grade. When it is not this it is often an improvement in method which would be suitable for children of all degrees of ability. It is worth while, in spite of the comparative lack of success up to the present, to investigate further the possibility of making genuinely qualitative adjustments.

The acceleration of the gifted child so that he enters high school and college early is often objected to on the ground of possible harm to the youth from associating at that level with those who are more mature. There is probably danger of maladjustment if the youth enters high school or college too young. Possibly two years below the usual age is the limit of safety in the ordinary case. But we must remember that there is a variation of as much as three or four years in physiological maturity, and also that intellectual equality constitutes part of the basis of association on a common

[1] For a fuller discussion of this method of adjustment see Frank N. Freeman, "The Treatment of the Gifted Child in the Light of the Scientific Evidence," *Elementary School Journal*, XXIV (May, 1924), 652–61.

footing. If the gifted child can be accelerated two years or so without suffering social maladjustment it is a decided advantage, for it is such children who are proper candidates for professional training, and the reduction of age of entrance upon the professions would be very beneficial.

The use of ability grouping as a mode of adjustment to individual differences has been and is still widely debated. The matter is discussed at length in the Thirty-fifth Yearbook of the National Society for the Study of Education, Part I, entitled, *The Grouping of Pupils.*[1] The chief objections are that, first, the basis of classification is not sufficiently accurate; second, that classification in one subject or field of work is not the same as that in other subjects; third, that there is too much of a dead level of ability, particularly in the lower group, and that it is difficult to handle the lower group; and, fourth, that it is undemocratic or that it induces an attitude of undue complacency in the bright children and discouragement in the dull children. It will be seen that there is some contradiction in these objections. The opinion is held, on the one hand, that there is not sufficient homogeneity in so-called homogeneous groups and not sufficient distinction between the groups and, on the other hand, that the distinction between the groups is too great and that there is too much dead level of ability within each group. It would appear that these objections cancel each other. So far as the effect on the personality of children is concerned, experience does not seem to indicate that it is serious. Furthermore, the contrast in ability of children seems to be more evident when those of wide ranges of ability are placed together than when they are placed in more or less homogeneous groups. We cannot

[1] *The Grouping of Pupils*, Thirty-fifth Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Co., 1936.

prevent children from recognizing differences in ability. It appears to be better to adjust to them than to disregard them.

It may be a question, however, whether homogeneous grouping is the best mode of adjustment. This question is particularly pertinent when corresponding modifications are not made in the content of the curriculum or in the demands upon the children of the various degrees of ability. If, for example, dull children are held to the same standard, the bright children will still mark time and the dull children will still be working at a disadvantage. The attempt to carry the dull children through the regular course will be a difficult and heartbreaking task. It seems necessary, therefore, to make a differentiation in requirements as well as to classify pupils according to ability.

If this differentiation in requirements is made, another mode of adjustment to individual differences may be better than routine homogeneous grouping. Such a plan has been in operation for some years in the mathematics departments of a number of high schools. After using homogeneous grouping for a number of years the difficulties which have been mentioned were recognized. Homogeneous grouping was then given up and an adjustment to individual differences was effected by making special provision for pupils at both ends of the scale of ability. The duller pupils were encouraged to take a course in which the practical applications of mathematics were stressed, such as general mathematics or useful mathematics, and the technical aspects of algebra and geometry were not included. The gifted pupils, on the other hand, who showed promise of high attainment and who might be expected to use mathematics in their professional careers as engineers or scientists, were permitted to enter upon the usual sophomore course in algebra in their freshman year. In this way an adjustment

was made to individual differences without a radical modification in the curriculum requirements. Similar adjustments may be made in this or other subjects to meet local requirements.

Another type of reorganization which is made for the purpose of adjusting the work of the school to differences in ability is individual instruction. Instead of treating the group of relatively homogeneous ability as a progress unit, each individual is treated as a distinct unit. The plan is ordinarily confined to the "tool" subjects such as reading, writing, arithmetic, and spelling. According to the view of the chief advocates of individual instruction, however, mental tests are not of much value in predicting or controlling rates of progress, since there is not much correlation between intelligence scores and rates of progress, or indeed between rates of progress in the various subjects themselves.[1] While this is probably an exaggeration, it is true that mental tests are of less use in individual instruction than in ability grouping. We shall therefore not discuss individual instruction further.

Mental tests should not be used as a sole basis either for determining the age of admission of the child to school, or of classifying into ability groups. The various other facts about the child which have already been mentioned should be taken into account. If a child is unusually large for his age, this fact should weigh in favor of promotion, or of classification with an advanced group. If he is unusually small, this fact should weigh in the opposite direction. Good

[1] See Carleton W. Washburne, "The Attainments of Gifted Children under Individual Instruction," *The Education of Gifted Children*, pp. 247–61. Twenty-third Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Co., 1924. See also *Adapting the Schools to Individual Differences*. Twenty-fourth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Illinois: Public School Publishing Co., 1925.

health, again, should weigh in favor of advancement, and poor health of holding back. Individual cases should be dealt with in the light of all the facts.

A word should be said about the procedure to be followed in giving and interpreting a test for the purpose of classification or of promotion. A rough or basic classification may be made by means of one or more group tests. Two group tests are more valid than a single test, and where it is possible it is advisable to give two tests and take both scores into consideration. If the scores agree, they indicate that the rating of the child is fairly reliable. If they disagree, it is necessary to secure additional evidence before coming to a decision. The same may be said concerning the relation of a child's score in an intelligence test to his standing in school work, in case a previous record is available. If the child's intelligence rating agrees with his educational rating, the intelligence rating thereby receives some confirmation. On the other hand, if the two disagree, the two ratings may or may not be reliable. It is, of course, possible that the child is bright and lazy, or suffers from some handicap which impairs his school achievement. Before concluding that this is the case, however, we should make sure that the intelligence rating is an accurate one. This means that additional tests should be given in order to confirm the rating of the first test. Similar confirmation is desirable when the rating of the test disagrees with the judgment of the teacher concerning the pupil's intelligence.

It is desirable, furthermore, to check up the rating of the intelligence test in the case of children at the upper or lower extreme, and possibly also in the case of children at the border line between two groups. It is particularly desirable to be sure of the rating of a child who receives a very low score. A low score is more likely to be due to an error than a high score. A child may fail to do himself justice because

he does not understand the directions, because he is in poor physical condition, because he is emotionally overwrought by the test, or by some experience just previous to the test.

5. The use of tests in selecting children for special classes

There are two types of special classes for which children may be selected on account of their intellectual capacity. The one is the class for backward children, and the other the class for gifted children. Such special classes usually differ from the homogeneous groups which have been mentioned in that they represent more extreme differences. Special classes for backward children include not only those who are somewhat slower in learning than the average, but those who are so defective that they cannot, even with more time, master the ordinary curriculum of the school. The group of gifted children represents those who can either proceed at a very accelerated pace, or perhaps require a radically different type of treatment from the majority of children. These special classes may contain from 5 per cent to 10 per cent of the children at either extreme.

The use of tests in selecting children for these special classes does not differ in principle from their use in selecting for other purposes. The chief difference is that somewhat greater care should be exercised, particularly in selecting children for the lower grade classes. In the majority of cases, to be sure, children at either extreme stand out more prominently and are more easily identified than those in the middle of the scale. The necessity of additional care rests, then, not upon the difficulty of identifying children, but upon the practical importance of avoiding mistakes in their selection. Because of the desirability of accuracy in selection, and because of the smaller numbers which are concerned, the final selection should probably be made upon the basis of individual examination.

6. The use of tests in educational guidance

When the pupil arrives at the point at which election among courses of study or subjects is possible, the advisor of a student may use intelligence tests to good advantage in his guidance. An illustration of the success of advice made with the help of tests is given in an experiment by Proctor.[1] Proctor gave intelligence tests to a group of eighth-grade pupils about to enter high school. The type of advice which was then given may be illustrated from two cases.

CARD NO. 3

Roe, Richard	Chronological age: 14 yrs., 4 mos.
Score Army Scale — 150	Stanford-Binet mental age: 16 yrs., 9 mos.
Army Scale mental age: 17 yrs., 4 mos.	
Army Scale I.Q. — 120	Stanford-Binet I.Q. — 117

. . . . . .

High school subjects which pupil desires to take: English, History, Algebra, French	Educational plans: To finish high school then attend a university or the U.S. Naval Academy. Vocational ambition: Chemical engineer or naval officer.

. . . . . .

Grade of work done in elementary and intermediate schools: Very poor. Estimated as "average" by some grade teachers, and as "below average" by others.

. . . . . .

Comment of examiner: Boy has ability but needs to be waked up. Suggest that he take general science in place of history for first year. Also suggest that he be placed in first division in algebra where he will have to work. He will need to develop ability in both science and mathematics if he is to follow his vocational ambition.

---

[1] W. M. Proctor, "The Use of Psychological Tests in the Educational Guidance of High-School Pupils," *Journal of Educational Research*, I (May, 1920), 369–81.

CARD NO. 4

Brown, Carrie
Score Army Scale — 100
Army Scale mental age: 14 yrs., 0 mo.
Army Scale I.Q. — 89

Chronological age: 15 yrs., 7 mos.
Stanford-Binet mental age: 14 yrs., 2 mos.
Stanford-Binet I.Q. — 90

. . . . . .

High school subjects which pupil desires to take:
English
Algebra
Latin
Typing
Drawing

Educational plans: Go to Mills College

Vocational ambition: To be a Chemist

. . . . . .

Grade of work done in intermediate and grammar schools: Grades in 8A class only fair, even in work that is being repeated. Estimates of elementary and intermediate teachers: "slow" but a conscientious worker.

. . . . . .

Comment of Examiner: Should be discouraged as to taking Latin. Algebra doubtful, but if she insists in view of desire to go to college assign to second division.

The experiment justified itself in the larger retention of pupils in high school and in the reduction of failures, as is shown by Table XXVII. Provided the precaution is taken of including other facts as determining factors besides the

TABLE XXVII. COMPARATIVE FACTS REGARDING "GUIDED" AND "UNGUIDED" GROUPS OF HIGH-SCHOOL PUPILS

GROUP	Out at work	Per cent	Out by transfer	Per cent	Failed 1 subject	Per cent	Failed 2 or more	Per cent
Guided....	1	4.5	2	9.0	4	18.0	0	0.0
Unguided..	13	12.0	14	13.0	33	31.0	11	10.0

intelligence tests alone, and the counsel is given in the form of advice rather than compulsion, there are very large possibilities in the use of intelligence tests for this purpose.

The next step in the counseling of pupils is to advise them regarding the selection of major courses of study, as contrasted with individual subjects. As the various subjects differ among themselves in the demand which they make, so the courses differ. It has been demonstrated that the so-called vocational or commercial courses in the high school demand less general intellectual capacity, or less of the capacity which is measured by our tests, than do the academic or college preparatory courses. This is shown by the fact that the students in the vocational courses have a considerably lower standing than those in the other courses. Furthermore, there is less correlation between the standing in commercial or vocational work and the scores on the tests than between the standing in academic work and the scores on these tests. The tests may be used in exactly the same way in advising pupils which course to take, then, as in advising them which subject to take.

If the tests of primary abilities or group factors which are now being experimented with by a number of psychologists prove to be valid methods of analyzing ability into its factors, such tests will be valuable adjuncts to the general intelligence scales in both educational and vocational guidance. If it proves to be possible to measure satisfactorily the distinctive kinds of ability necessary in the different school subjects and in the different vocations, guidance may be based specifically on the outcome of the application of these tests. Sufficient evidence has not yet been accumulated to indicate whether this can be done. It is clear that somewhat different abilities are required in the various academic and vocational pursuits. The differences were neglected by the earlier psychologists who over-

emphasized the general character of intelligence and the share which intelligence has in the attainment in various pursuits. Whether general intelligence exists as a psychological factor or whether it is merely a composite of primary abilities is still a moot question. This question will be discussed in the chapter on "The Nature of Ability." In the meantime, it may be said that there are sufficient distinctions among the abilities in particular lines to make it desirable to take them into account in advising the pupil what courses to pursue.

Another phase of educational guidance has to do with the length of time a child shall remain in school. There is, in fact, as will be shown in Chapter XV, a close correlation between the amount of schooling a child receives and his intelligence, or between the age of dropping out of school and intelligence. The school at present acts in a measure as a selective agency. The brighter pupils remain longer, and the duller ones drop out sooner. This correlation is undoubtedly due in part to the fact that the larger amount of schooling causes a higher score in the mental test. After we have made due allowance for this fact, however, there remains a considerable degree of correspondence due to the fact that the higher levels of the school demand more intelligence than the lower ones.

This fact has been taken by some as a criticism of the school. The school, according to these critics, should furnish a type of training at every level up to and including the college which is adapted to every range of ability. This would involve the addition to the types of training given in the college, and even in the high school, of work which is more largely manual in nature and which demands less abstract thinking. How far it is the function of the high school and the college to add to the present type of work courses of this character is a matter of broad educational

policy which is not our problem here to decide. It is the opinion of the writer that there is a distinct limitation upon the desirable extension of full time high-school and college work in this direction. It is probably desirable that there should be a very large extension of part time and continuation work, extending even to adult education, in order that those who are not fitted to continue indefinitely the full work of these higher institutions may add to their elementary school training further education suited to their capacity, so as to fit them to perform the duties of citizenship and to develop habits of making a wholesome use of their opportunities for recreation. Continuation training should also serve to improve vocational fitness, including home making on the part of women.

If this conception of the function of secondary and higher education is accepted, a distinction will have to be made between individuals, based upon their fitness to continue their education to the higher levels of the high school and college. Even if the high school and college should ultimately be reorganized so as to provide full time training to suit the capacities of everybody, they are not so organized at the present time and it is now necessary to make a distinction. Some individuals are capable of continuing through high school and through college, while others are not. In addition to the evidence from the correlation between the amount of schooling and intelligence we may cite illustrative findings on the relation between intelligence tests and the probability of success in high school and college. Terman estimates that an I.Q. of 90 is necessary for successful high-school work, and an I.Q. of 100 for successful college work. Standards of this sort may be taken for general guidance, and a student may be advised as in the case of the selection of subjects or of courses. In all cases, of course, the character of the student's previous work in school is to

be taken into account as well as his intelligence test score.

In the case of the college student, the question arises whether he shall continue beyond college into the professional school, and if so, which profession he shall select. It is coming to be more commonly necessary to make this decision before the end of the college course. At least it is advantageous to the student to do so, because of the fact that he can, in his last two years of college, take courses which are preparatory to his professional studies. It is not clear that professional schools as a whole require a higher degree of intelligence than does the liberal arts college. In

TABLE XXVIII. STANDING IN THE ARMY ALPHA TEST IN DIFFERENT SCHOOLS WITHIN A UNIVERSITY

UNIVERSITY	SCHOOL OR DEPARTMENT	MEDIAN SCORE
Ohio State University[1]	Arts, Commerce and Journalism	147
	Medicine	142
	Law	142
	Engineering	141
	Education	137
	Agriculture	133
	Arts	133
	Pharmacy	125
	Dentistry	115
	Veterinary Medicine	112
University of Illinois[2]	Graduate School	154
	Literature, Arts and Sciences	145
	Engineering	144
	Commerce	143
	Agriculture	139

[1] E. L. Noble and G. F. Arps, "University Students' Intelligence Ratings According to the Army Alpha Test," *School and Society*, XI (Feb. 21, 1920), 233–37.

[2] Yoakum and Yerkes, *Army Mental Tests*, p. 17. New York: Henry Holt & Co., 1920.

fact, the students in certain professional departments appear to have a lower average standing than do the liberal arts college students. It may be that the chief form of guidance at the college level should consist in advising students which type of professional preparation to select, in case they wish to go forward into a professional school. The data which have been gathered in two universities and which are shown in Table XXVIII seem to give some color to this suggestion.

7. The use of tests in maintaining the adjustment of a pupil to his work

The necessity of adjustment is most evident in the case of the pupil who fails in one or more of his subjects. Tests have been used in such cases to assist in the analysis of the failure and in the determination of the type of the remedial work or other treatment which is necessary to overcome the failure and to prevent its return. If the pupil's failure is due to lack of ability, this will be revealed by the test. If the test does not show lack of ability, the cause must be looked for elsewhere. In some cases, as, for example, in reading, the child may possess sufficient general ability, but may have a special disability. In such cases it is necessary to apply special ability tests as far as they may be available. Special ability tests are most highly developed in the field of music. In the case of other subjects it is necessary to determine whether the child possesses special disability by means of tests of achievement in the subjects themselves. If the child's general ability is normal, but he fails in particular subjects, we should first exhaust the possibility that this failure is due to lack of interest or to ineffective training at some point in the child's previous school career. Only when the failure cannot be explained on one of these grounds should we resort to the explanation of a special disability. Bronner discusses this whole question of special ability and

disability, and suggests the possible use of tests to discover them.[1] The teacher or the supervisor will not find that our present tests, however, are of very much value for this purpose.

If the cause of the failure has been determined to be lack of previous training or a special disability, remedial treatment is required. In the first case the remedial treatment is comparatively simple, since it merely consists of giving the child the training which he has missed in his previous career, or which has been inadequately given. Nothing new in the nature of training is necessary. If a special disability is to be overcome, on the other hand, considerable ingenuity must be exercised in order to find the type of training which will meet the peculiar necessities of the case. The subject in which the largest amount of work in the training of children with special disabilities has been done is reading. For a further account of the methods to be used the reader may be referred to the monograph on the subject by W. S. Gray.[2]

Failure must always be interpreted in a comparative sense. The child who receives passing grades in his work may be failing as much as the child who receives a grade below passing. Failure is to be considered in relation to the pupil's capacity. The very bright pupil who is doing mediocre work is failing in so far as his achievement falls below his capacity. Mental tests may therefore be used to stimulate brighter pupils to work up to their capacity. In some cases, pupils present problems in conduct because the work

[1] Augusta F. Bronner, *The Psychology of Special Abilities and Disabilities.* Boston: Little, Brown & Co., 1917.

[2] William Scott Gray, *Remedial Cases in Reading.* Supplementary Educational Monographs, No. 22. Chicago: Department of Education, University of Chicago, 1922.

See also *The Teaching of Reading: A Second Report.* Thirty-sixth Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Co., 1937; and Marion Monroe, *Children Who Cannot Read.* Chicago: University of Chicago Press, 1932.

which they are doing is so far below their capacity that it does not enlist their interest or engage their energy. Such pupils have sometimes become well behaved by being promoted and given work which was more nearly commensurate with their capacity.

If this use of mental tests to stimulate the bright pupils is to be successful in the long run, the pupil's willing coöperation must be secured and the work of the school must not be regarded by him as a task. The opportunity to do a high grade of work must be looked upon by him rather as a privilege than as an urgent requirement. Otherwise there is danger that, as the pupils become familiar with the uses which are made of tests, they will malinger, and the test score will fail to represent their true capacity.

8. The selection of applicants for college or professional school

As our educational system is now organized, admission is not restricted below the level of the college or the professional school except in the requirement of graduation from the next lower grade. In the college experiments have been made in recent years for the purpose of gathering information to indicate whether intelligence tests may be used as an appropriate means for selecting candidates for admission. Tests are not yet widely used for this purpose, although they are applied in a number of institutions as a matter of record and for purposes of later administrative dealing with the student.

Chapter XV

INTERPRETATION OF MENTAL TESTS

In the first chapter certain questions were raised concerning the meaning and interpretation of mental tests. It was pointed out that there are rather wide differences of opinion concerning the significance of tests. While these differences of opinion are more extreme among persons who are not well acquainted with tests, and have not used them, than among technical psychologists, nevertheless there is some divergence also in the opinion of psychologists themselves. In the course of the treatment of the development of mental tests, of their technique, and of their application in various fields, a good many facts have been presented which bear upon these problems and which will enable us to arrive at at least an approximate answer to them. At the end of the book, therefore, we shall attempt to bring together threads of facts and interpretation which have run through the discussion of the various topics, and weave them into a pattern which shall give a summary of the conclusions that seem justified in the present stage of the science.

1. Two fundamental problems in interpretation

Of the problems in the interpretation of mental tests which have attracted the attention of psychologists, two stand out as of wide significance. The first has aroused animated discussion in lay as well as in professional circles. Expressed in simple terms, it is this: Do mental tests measure native capacity, as they purport to do, or do they merely measure education and experience? The second question is somewhat more abstruse, but is still of practical

importance. It is concerned with the nature of ability — with the constitution of the ability, whether native or acquired, which is measured by the mental tests. The first question will be treated in this chapter, and the second in the next chapter.

It must be frankly recognized, and it is recognized by competent psychologists, that our mental tests are subject to definite limitations. The first limitation is that they measure intellectual capacity indirectly rather than directly. This must always be the case, since capacity is simply potentiality for behavior. A person exhibits capacity only as he acts, and it is only his acts which we can measure.

The types of limitation which arise from this fact are two. In the first place, we do not measure and cannot measure all of behavior. We are restricted to the measurement of particular samples of behavior. We must assume that these samples which we measure have been so selected that they constitute fair representatives of the individual's behavior as a whole.

Within certain limits this assumption is more than an assumption, since it is subject to statistical inquiry. We may determine statistically how many samples of reaction of the kind which are measured in the test are necessary to be secured in order that the test as a whole may be as reliable as it is possible to make it. We may determine, for example, whether five of the tests of the kind which are ordinarily used in our group point scales are as reliable as are ten, or if not, just how many are necessary in order that the maximum of reliability may be reached. But the question still remains whether our mental tests as they are ordinarily constituted include all the range of types of behavior which are necessary in order that we may have a complete sampling of those forms of reaction which constitute intellectual activity.

The second consequence of the fact that mental tests are indirect measures is that the behavior which they measure is conditioned not only by native endowment, but by experience and training. As a consequence, the measures which we secure by means of these tests are measures not merely of native endowment, but also of the results of training, or education. That this is true as a general principle nobody has recognized more clearly than have the psychologists themselves, and they have repeatedly pointed it out in their writings. The attacks of popular writers upon intelligence testing, which are based upon the contentions that tests are in part measures of training, are therefore directed against a straw man. But the psychologists, unlike the popular writers, are not content with recognizing this general principle that behavior is conditioned by both endowment and training; they endeavor further to analyze the facts and to determine as precisely as possible what bearings the facts have upon our interpretation of intelligence tests.

While it is true that psychologists generally recognize that the measurement of mental ability is indirect and that the performance of an individual on a test always involves in some measure the result of training and experience, there is a considerable difference in the importance which various psychologists attach to these factors. Some believe that the individual's ability, whether expressed in the performance on mental tests or in other forms of activity, represents in the main the individual's native endowment. Others believe that what the individual is, his abilities, and his performance are influenced to a much greater extent by the forces of environment and by the experience and activities which he has undergone throughout his life. This difference of opinion is represented by the estimate of the relative influence of heredity and environment or of nature

and nurture upon performance, including performance in mental tests.

We ordinarily think of this problem in the terms that have been used in stating it, that is, we think of the problem as concerned with the relative share of nature and nurture in the make-up of an individual. Our scientific approach to the problem, however, is directed toward a somewhat different question. We may ultimately apply the results of our study to the interpretation of the factors in the individual, but our investigation is directed to the explanation of the differences between persons as they appear in their behavior. It is only through the investigation of these differences that we are able to form any estimate of the factors in the make-up of an individual. The question, then, is put in this form: What is responsible for the fact that A is superior to B in general ability or in some special ability? Is it because he inherited a different constitution or because he has received a different training and has had different experiences?

Many psychologists have been predisposed to answer this question on the basis of somewhat indirect or presumptive evidence rather than on the basis of a direct investigation of the factors in individual differences. This bias had led many to conclude that mental ability, as measured by the tests, is almost entirely inherited. Thus they examine the direct evidence with a prejudice which predisposes them to the hereditary interpretation. These two lines of indirect evidence are the constancy of the I.Q. and the general conception of laws of inheritance.

The early users of intelligence tests were impressed with the fact that there is a good deal of consistency in the scores and the I.Q.s of pupils when tests are repeated. This consistency came to be called "the constancy of the I.Q." Because the I.Q. was found to be fairly constant the

conclusion was drawn that it represented fixed and unalterable native capacity.

This interpretation of the constancy of the I.Q. went beyond what was warranted by the facts. In the first place, the I.Q. is not completely constant. There is a certain amount of variation, and, as we have already seen in the chapter on "The Educational Uses of Tests," this variation is considerable and becomes greater with the passage of time. In other words, the I.Q. is not completely constant, only relatively constant. Even if the average change in I.Q. on repetition of a test were only five points after several years' testing as well as after a short interval of time, it would mean that half the changes are greater than five points, and some would be ten, fifteen, or more points. These larger changes need to be explained as well as the comparative constancy of the I.Q.s of children in which there is not a great change. In the second place, even if the I.Q. were even more constant than it is, it would not prove that this constancy comes from inherited capacity. It might only indicate that the environment is constant. We should expect the I.Q. to change on the hypothesis that it is affected by the environment only when there is a radical change in the environment itself. The environment of most children does not change radically during the period when they are in school and, therefore, there is no reason to expect a radical change in I.Q. Finally, the mental tests from which the I.Q.s are derived are ordinarily given after the child has had six or more years of training and experience during the most formative part of his life. If we assume that the I.Q. of the child who is tested for the first time at eight years of age is the joint product of his native capacity and his training, we should expect that I.Q. to be relatively stable unless he is subjected to great changes in environment. The constancy of the I.Q., therefore, does not prove much

on either side of the argument. It has been given much greater weight in the thinking of many psychologists on this problem than it deserves.

The biological facts of heredity have also been given undue weight by some psychologists in their interpretation of mental tests. No enlightened person questions the facts of heredity. He may, however, question some of the interpretations of these facts. It is a generally accepted fact, for example, that the original constitution of the individual, that is, his body, is determined by the genes which reside in the chromosomes of the germ cells (the ovum and the spermatozoon) of the parents. No one questions, further, that there are wide individual differences in the constitutions of different individuals as determined by the combination of genetic factors arising from the parental strains of these individuals. There may be differences of opinion concerning the amount of these differences, but nearly all psychologists would agree that they are large.

The split in opinion comes at the point of explaining how the individual develops and how the individual differences of mature individuals are to be explained. The one school believes that the constitutional differences which are inherited determine almost solely the individual's course of development and his ultimate characteristics. The other school believes that the outcome of the native constitution will be affected to a large degree by the surroundings, physical, social or intellectual, in which the individual is brought up. Both sides would agree that original constitution is inherited. The hereditarians would also say that adult characteristics are inherited. The environmentalists, on the other hand, would maintain that adult characteristics are the joint product of native constitution and of the influences of the environment. This point of view is very clearly brought out by the eminent biologist, H. S. Jen-

nings.[1] Jennings shows clearly that marked modifications in the physical characteristics of the mature individual may be brought out by changing the physical surroundings of the organism during the embryological period. He concludes from these facts that the laws of heredity do not enable one to say how far the adult individual is what he is because of genetic factors or because of environment. The relative share of these two sets of factors in making him what he is is a matter for investigation rather than a matter for inference from the biological laws of heredity. We should approach the problem, not with a presupposition in one direction or the other, but with an open mind to consider and interpret fairly the facts as we find them.

We may make an approach to the problem by various methods. The first method is a statistical one, and is called the method of partial correlation and multiple correlation. The partial correlation method may be illustrated in this way. Suppose we are finding the correlation between intelligence test scores and school achievement, and our group contains children of several different ages. As was pointed out in Chapter III, the fact that both test scores and school achievement are correlated with age will raise the correlation between these two factors. By using an appropriate formula the effect of age can be eliminated, so that we can find what the correlation would be if we had a group all of the same age.

Burt [2] has attempted to apply this method to the Binet-Simon scale. He wishes, among other things, to determine how far the correlation between the intelligence test scores and school achievement is due to the fact that both are

[1] H. S. Jennings, *The Biological Basis of Human Nature.* New York: W. W. Norton & Co., Inc., 1930. See also H. S. Jennings, *Prometheus.* New York: E. P. Dutton & Co., 1927.

[2] Cyril Burt, *Mental and Scholastic Tests.* London: P. S. King & Son, Ltd., 1922.

determined by native capacity, and by so doing to determine how far this correlation is due to other circumstances, such as the inclusion in the test of scholastic material. The reader will perceive that this requires, by hypothesis, a true measure of native intelligence — one which is independent of schooling.[1] Burt takes as such a measure his reasoning test. This is regarded as a measure of native intelligence, because it does not correlate with the standing of children in school tests. This conclusion has two serious difficulties. The first is that such a criterion does not give us a test which is *independent of the amount of schooling a child has received*, which is what we want, but rather of his attainment in school, which is a very different matter. The second is that a test which has no correlation with school achievement cannot be a measure of intelligence as we understand it. Intelligence is certainly manifested in part by superior school achievement.

The independent measure which is needed is one which does not correlate with the amount of schooling one has had, on the one hand, but does give positive evidence of measuring intelligence — as indicated, among other things, by scholastic attainment — on the other hand. That, of course, is just what we are looking for in our intelligence tests. When we get it we can test our present intelligence tests by means of it, but the partial correlation method will not give it to us, because we need it before we can calculate the partial correlation.

Because of the rather wide notice it has received, and because it appears to give in an accurate formula the constituent factors in the Binet-Simon scale, we may note

[1] Karl J. Holzinger and Frank N. Freeman, "The Interpretation of Burt's Regression Equation," *Journal of Educational Psychology*, XVI (December, 1925), 577–82. The intricacies of the question, including the ambiguity of the term *schooling*, cannot be gone into here.

Burt's next step. By using a number of partial correlations between the various factors of Binet score, Schooling (school achievement — not amount of schooling), Intelligence (Burt Reasoning score) and Age, he calculates a multiple correlation which purports to show the relative share of the factors in the Binet score. The formula is as follows (p. 180):

$$B = .54\,S + .331\,I + .11\,A$$

Where B = Binet-Simon score
S = School achievement
I = Burt Reasoning score
A = Age

In interpreting this formula Burt says, "In determining the child's performance in the Binet-Simon scale, intelligence can bestow but little more than half the share of school, and age but one third the share of intelligence."

To further inquire into the validity of this interpretation Holzinger has taken Burt's data and calculated the equations with each of the factors on the left side. Its *reductio ad absurdum* appears in the formula for age:

$$A = .15\,B + .51\,S + .03\,I$$

According to Burt's interpretation this means that, in determining the child's age, Binet score bestows somewhat less than a third the share of schooling. Or, to use another expression which he employs, a child's age is a measure not only of the Binet score, but largely if not mainly of "the mass of scholastic information and skill which in virtue of attendance more or less regular, by dint of instruction more or less effective, he has progressively accumulated in school" (p. 182). Thomson [1] criticizes such an interpretation

[1] Godfrey H. Thomson, "The Interpretation of Burt's Regression Equation," *Journal of Educational Psychology*, XVII (May, 1926), 301–08. See also Karl J. Holzinger and Frank N. Freeman, "Rejoinder on Burt's Regression Equation," *Journal of Educational Psychology*, XVII (September, 1926), 384–86.

and says: "... the most probable hypothesis here is simply that school work (degree of attainment or amount of schooling) assists the child to answer Binet questions, and a very heroic pedantry is required by anyone who would refrain entirely from saying so." If this is the case, many psychologists achieve heroic pedantry, for they maintain that answering Binet questions depends on the possession of native intelligence and that it is this same intelligence which so largely determines success in school. Statistically, one interpretation is as acceptable as the other.

The method of partial correlation, then, as a means of investigating the relative share of nature and nurture, is subject to two difficulties. In the first place, the mere fact of the association between two factors does not indicate which is cause and which is effect. We may interpret either one as being the cause of the other according to our common sense or according to our interpretation of other lines of evidence. The ultimate explanation, then, goes back to other evidence. In the second place, in order to determine the relation between two factors with a third constant it is necessary to have independent measures of these two factors. That is, if we wish to determine the effect of hereditary constitution on achievement with education constant we have to have both a measure of native constitution and a measure of education. Or, on the other side, if we wish to measure the relation of achievement to education with heredity constant we need to have measures of both of these two factors. Since our main problem is to determine how far mental tests are determined by heredity and how far by environment, it is apparent that we must have the answer to the question before we can investigate it by this method. We must, therefore, assume our answer and our argument is in a circle, that is, we beg the question.

A favorite method of inquiring into the relative share of

native constitution and environment is to compare various groups of persons. If we find one group superior to another we interpret this superiority as due either to heredity or environment. Many such group comparisons have been made between vocational groups, geographical groups, racial groups, and so on. The singular thing about these comparisons is that they may be interpreted as indicating the effect of hereditary differences by those whose predilection leans toward this explanation or they may be explained as illustrations of environmental differences by those who prefer this explanation. Since they are susceptible to both types of interpretation it is clear that they suffer from a fundamental defect. We may illustrate some of the group comparisons and the interpretations which have been made of them and then point out the defects from which they suffer.

2. Differences between vocational groups

We may begin with a comparison between men in different occupations.

It is evident that if we take the test scores at their face value there is a vast difference between the average capacity of men who are engaged in different occupations.

It may be that this difference is due partly to the fact that some occupations fit men directly to do well in intelligence tests and others do not. If the differences are wholly due to a difference in the effect of the occupational activity itself, we should not expect to find that the children of men in the various occupations should exhibit like differences in intelligence test scores. The studies which have been made indicate, however, that differences of a similar sort exist among the children as among the parental groups. For

TABLE XXIX. THE MEDIAN SCORES IN ARMY ALPHA OF MEN CLASSIFIED AS BELONGING IN CERTAIN OCCUPATIONS*

OCCUPATION	NO. OF CASES	MEDIAN SCORE
Farmer	6886	48.3
General machinist	1251	62.8
Railroad clerk	308	91.4
Bookkeeper	458	100.9
Accountant	202	117.9
Stenographer or typist	402	115.0
Mechanical engineer	45	109.7
Civil engineer	53	116.8

* Robert M. Yerkes (Editor), *Psychological Examining in the United States Army*, pp. 824–29.

example, Pressey and Ralston [1] tested 548 children and calculated the percentage of each of four occupational groups which tested above the median. They are as follows:

OCCUPATION OF FATHERS	PERCENTAGE ABOVE MEDIAN
Professional	85
Executive	68
Artisan	41
Laborer	39

Other studies have yielded similar results.

An intensive study of the relation of occupation to the intelligence of children appears in Terman's investigation of gifted children. Terman first selected a group of children,

[1] S. L. Pressey and R. Ralston, "The Relation of the General Intelligence of School Children to the Occupation of Their Fathers," *Journal of Applied Psychology*, III (1919), 366–73.

all of whom had an I.Q. of about 140 or above, and then tabulated the occupations of their fathers and compared the distribution with the distribution of occupations in the population in general. The summary result of this comparison is shown [1] in Table XXX.

TABLE XXX. OCCUPATION OF 560 FATHERS OF GIFTED CHILDREN CLASSIFIED ACCORDING TO THE CENSUS REPORT

GROUPS	Proportion among fathers of gifted children (per cent)	Proportion in population of Los Angeles and San Francisco (per cent)	Per cent of quota among fathers of gifted children (per cent)
Professional group.....	29.1	2.9	1003
Public service group.....	4.5	3.3	137
Commercial group.....	46.2	36.1	128
Industrial group.....	20.2	57.7	35

An intensive study of the relation of the intelligence of children to the occupation of their parents is reported by Byrns and Henmon.[2] These investigators tabulated the scores of over one hundred thousand high-school seniors of the State of Wisconsin. They found an enormous difference between the children of men of different occupations. However, there was great overlapping between the groups and hence the correlation between the occupations and the intelligence of the children was only .18. Because of this large overlapping more than 50 per cent of the children whose scores were above the median were from the two lowest occupational groups.

[1] Lewis M. Terman and Others, *Mental and Physical Traits of a Thousand Gifted Children*, p. 63. Genetic Studies of Genius, Vol. I. Stanford University, California: Stanford University Press, 1925.

[2] Ruth Byrns and V. A. C. Henmon, "Parental Occupation and Mental Ability," *Journal of Educational Psychology*, XXVII (April, 1936), 284–91.

Terman also rated the occupations of the fathers of the gifted children according to the level of intelligence which they required, as measured by the Barr scale. The distribution of their ratings in comparison with that of the population as a whole is shown in Table XXXI.

TABLE XXXI. DISTRIBUTION OF RATINGS FOR FATHERS OF 526 GIFTED CHILDREN AND FOR THE GENERAL ADULT MALE POPULATION ACCORDING TO THE SCALE FOR RATING THE INTELLIGENCE OF OCCUPATIONS

(Devised by F. E. Barr, p. 72)

RATING	FATHERS OF GIFTED (per cent)	ADULT MALES IN GENERAL
15 or above	26.8	2.2
12–15	26.8	4.5
9–12	36.1	37.0
6– 9	8.9	13.4
3– 6	1.3	42.9

Taking into consideration these various lines of evidence, it is clear that persons in certain occupational groups have higher test ratings than do members of other occupational groups, that the ratings of the children of the families in occupational groups differ as do their parents, and that the occupational scale runs from the professions at one end to common labor at the other. It is evident that these occupational differences may be attributed to either of the two causes. The occupations may be considered as the selective agencies in that the professions attract individuals of high native intelligence and the simpler occupations those of lower native intelligence. The same explanation may be

applied to the children. On the other hand, it may be supposed that the preparation for some occupations and the actual performance of the duties of these occupations not only demand but cultivate higher intelligence than do others. It may also be supposed that the children of men in some occupations have greater advantages than those of men in other occupations. The evidence from these differences, therefore, is not conclusive.

3. Differences between geographical groups

There is ample evidence that wide differences exist in the test scores of persons living in different localities. Perhaps the most striking of these are the differences between the standing in the army tests of recruits from the various States. In the *Army Report* is given the distribution of the

TABLE XXXII. THE MEDIANS OF THE STATE DISTRIBUTIONS AS CALCULATED FROM TABLE 200 OF THE ARMY REPORT

STATE	STATE	STATE	STATE
Ore. 79.9	Ohio 67.3	R.I. 62.9	Tex. 50.9
Wash. 79.2	Me. 67.0	N.H. 61.9	N.J. 48.7
Calif. 78.1	Neb. 66.2	Mo. 59.5	S.C. 47.4
Conn. 73.6	Pa. 65.1	S.D. 58.3	Tenn. 47.2
Idaho 73.5	N.Y. 64.5	N.D. 57.1	Ala. 46.3
Utah 72.2	Iowa 64.4	Wis. 56.5	La. 45.2
Mass. 71.6	Minn. 64.0	Va. 56.3	N.C. 43.2
Colo. 69.7	Kan. 63.9	Md. 56.2	Ga. 42.2
Mont. 68.5	Ill. 63.8	Ind. 56.1	Ark. 41.6
Vt. 67.5	Mich. 63.3	Okla. 52.5	Ky. 41.5
			Miss. 41.2

Alpha scores of 40,530 men (whites) classified by the states of their residence. The medians of these distributions (excluding the states for which there are fewer than 500 cases) have been calculated by Alexander,[1] as shown in Table XXXII. They have been taken from Table 200 of the *Army Report.* Such startling differences, based upon such careful and extensive measurements, indicate the presence of some factor or combination of factors of great magnitude.

In Table XXXIII are shown the distributions of the ratings of recruits in five Northern and five Southern states.

TABLE XXXIII. THE PERCENTAGE DISTRIBUTION OF LETTER GRADES OF WHITES IN TEN CAMPS

STATE	CAMP	NUMBER	LETTER GRADES		
			D and D −	C −, C and C +	A and B
Ill.	Grant	7,671	18.2	69.6	12.2
Kan.	Funston	6,058	16.1	71.2	12.7
Mass.	Devens	8,247	20.1	63.5	16.7
Mich.	Custer	4,933	20.6	67.1	12.5
N.Y.	Upton	7,876	22.1	66.6	11.3
Average.....			19.5	67.3	13.2
N.C.	Wadsworth	8,243	27.0	61.2	11.8
Ga.	Gordon	4,503	31.8	60.9	7.3
Tex.	Travis	6,514	34.4	56.4	9.5
Ind.	Meade	4,638	37.4	53.8	8.7
Va.	Lee	3,512	42.7	51.7	5.8
Average.....			33.3	57.5	9.2

[1] Herbert B. Alexander, "A Comparison of the Ranks of American States in Army Alpha and in Social-Economic Status," *School and Society,* XVI (September 30, 1922), 388–92.

Similar differences appear when we compare pupils in cities with pupils in small towns or in the country. Since all the studies with which the writer is familiar show the same differences, two illustrations will suffice. Pressey and Thomas [1] made a comparative study of 2800 city children and 538 country children. They express the results in terms of the percentage of the country children who excel the median of the city children. The amounts are as follows:

PERCENTAGE OF COUNTRY CHILDREN MAKING SCORES ABOVE THE MEDIAN OF THE CITY CHILDREN

	AGE			
	10	11	12	13
Per cent.....	29	33	21	25

Book [2] made a similar comparison of high-school seniors, based on an extensive survey. His summary of the comparison between the high-school seniors in city and rural schools is given in Table XXXIV. While the differences are not so great as are those which have been reported for elementary-school children, they still persist. The lesser difference may indicate that at least part of the superiority of city children may be due to superior training, since it becomes less as the amount of training increases. In confirmation of this suggestion, Superintendent R. H. Bracewell, of Burlington, Iowa, reports, in an unpublished study, that the superiority in tests of city pupils on entering high school is reduced by the beginning of the sophomore year.

In harmony with the differences between whole sections

[1] S. L. Pressey and J. B. Thomas, "A Study of Country Children in (1) a Good and (2) a Poor Farming District, by Means of a Group Scale of Intelligence," *Journal of Applied Psychology*, III (1919), 283–86.

[2] William F. Book, *The Intelligence of High School Seniors*. New York: Macmillan Co., 1922.

TABLE XXXIV. PER CENT OF SENIORS FROM CITY AND RURAL HIGH SCHOOLS SCORING AT VARIOUS INTELLIGENCE LEVELS — BASED ON ABOUT 2400 CASES

SECTION OF STATE	TYPE OF SCHOOL	PER CENT ABOVE MEDIAN	MEDIAN SCORE
Northern	City	60	141
	Rural	43	134
Central	City	58	141
	Rural	46	135
Southern	City	49	136
	Rural	36	130

of the country and between the city and the rural district, marked contrasts have been found between the more and less favored parts of the same city. Yerkes and Anderson,[1] for example, compared the point scores of young children in two city schools of Cambridge, Massachusetts, "which differed radically in the social and economic status of their pupils." There were 54 individuals in each group and matched pairs were selected who were approximately equal in age. The average score of the favored boys was 37.2 and of the unfavored boys 29.5. The average score of the favored girls was 41.0 and of the unfavored girls 32.6. The difference in each case is about 20 per cent.

A somewhat more extensive comparison of children in a favored district and children in a mill district of Columbia, South Carolina, was made by Strong.[2] The results, which

[1] Robert M. Yerkes and Helen M. Anderson, "The Importance of Social Status as Indicated by the Results of the Point-Scale Method of Measuring Mental Capacity," *Journal of Educational Psychology*, VI (March, 1915), 137–50.

[2] Alice C. Strong, "Three Hundred Fifty White and Colored Children Measured by the Binet-Simon Measuring Scale of Intelligence: A Comparative Study," *Pedagogical Seminary*, XX (December, 1913), 485–515.

are summarized in Table XXXV, also include a comparison of the scores of negro children. The Binet-Simon scale was used.

TABLE XXXV. DISTRIBUTION OF THE RATING OF THREE GROUPS OF CHILDREN TESTED BY THE BINET SCALE

RATING	FAVORED WHITES		UNFAVORED WHITES		COLORED	
	No.	%	No.	%	No.	%
More than 1 year backward.....	5	5.3	11	18.3	21	25.6
Satisfactory.....	80	84.2	49	81.6	61	74.4
More than 1 year advanced.....	10	10.4		0		0

Our final comparison between groups which are classified according to place of residence is between immigrants from the various European countries. This comparison is based on the Army tests, and is worked over in terms of a combined scale by Brigham.[1] The average scores of men coming from the various countries are as shown in Table XXXVI.

TABLE XXXVI. SCORES OF VARIOUS IMMIGRANT GROUPS IN THE ARMY SCALE

England.....	14.87	Belgium.....	12.79
Scotland.....	14.34	Ireland.....	12.32
Holland.....	14.32	Austria.....	12.27
Germany.....	13.88	Turkey.....	12.02
Denmark.....	13.69	Greece.....	11.90
Canada.....	13.66	Russia.....	11.34
Sweden.....	13.30	Italy.....	11.01
Norway.....	12.98	Poland.....	10.74

In common with the other comparisons between environmental groups these differences are susceptible to more than one possible explanation. The proponents of Nordic race

[1] Carl C. Brigham, *A Study of American Intelligence*, pp. 120–21.

superiority hold, first, that these various nationals can be classified according to race into three groups, second, that the national groups of immigrants are fair representatives of their countrymen in general, and, third, that the differences in the test scores are unqualified measures of native intelligence. On the first point there is dispute among anthropologists. On the second point we have little or no knowledge. On the question whether the test scores are affected by other factors than native capacity we have two further comparisons which at least raise a doubt.

The first of these comparisons was made by Brigham. He gives the average scores of immigrants who have been in the United States for different periods of time. The scores by five-year periods are as follows: [1]

0–5 years	6–10 years	11–15 years	16–20 years	Over 20 years
11.41	11.74	12.47	13.55	13.82

On the racial difference hypothesis this might be due to an increase in proportion of immigrants from the alleged lower racial stocks in recent years, but a comparison of the proportion coming in during the first and the second decade of the nineteenth century indicates that this is not the case.[2] The racial advocate is then forced to suppose that immigrants are being drawn from progressively lower strata of the countries from which they come. This supposition seems rather strained in comparison with the simple hypothesis that being under the influence of American schooling and environment for periods varying from five to twenty-five years enables men to make a higher score than they otherwise would make.

This hypothesis is somewhat strengthened by the second

[1] Brigham, *op. cit.*, p. 89.

[2] Brigham, *op. cit.*, p. 163.

fact, which is stressed by Bagley.[1] He found that there is a correlation between the ratio of elementary-school enrolment to the populations of foreign countries from which immigrants come and the average test scores, the coefficient being .91. Before venturing upon further interpretation let us return to the inspection of other group differences.

4. Differences between racial groups

We turn to a comparison of the test scores of racial groups which are fairly clearly marked. The largest scale comparison on record is that between negroes and whites in the army. A comparison of the letter distribution of scores in Army Alpha between the whites and negroes of five Northern and five Southern camps is given in Table XXXVII.[2]

TABLE XXXVII. COMPARISON OF THE SCORES OF WHITES AND NEGROES IN ARMY ALPHA

COMPOSITION OF GROUP	SCORE IN LETTER GRADES		
	D and D− (per cent)	C− C and C+ (per cent)	A and B (per cent)
Whites, five Northern camps.....	19.4	67.6	13.1
Negroes, five Northern camps.....	45.3	51.1	3.6
Whites, five Southern camps.....	34.8	56.8	8.6
Negroes, five Southern camps.....	78.7	20.6	.7

This table shows that over twice as many negroes make low scores as are made by whites of the same region of the country, whereas the preponderance of whites over negroes

[1] W. C. Bagley, "Army Tests and the Pro-Nordic Propaganda," *Educational Review*, LXVII (1924), 179–87.

[2] Robert M. Yerkes (Editor), *Psychological Examining in the United States Army*, pp. 679, 719.

making high scores is still greater. The table also illustrates further the great difference in the scores of Northern and Southern men. Since this difference cannot be attributed to race there must be some other factor besides race to account for the superiority of some groups over others. If this factor consists in whole or in part in education or some other environmental influence, the difference between negroes and whites may be due in part to environment.

A similar difference in favor of whites is found in a comparison of white and negro children. In Strong's study, summarized in Table XXXV, a group of negro children were found to stand below the white children of the mill district, and much below the white children of the favored district. A marked difference in both a verbal and a nonverbal group test was found by Sunne.[1] She gave the Myers Mental Measure, a nonverbal test, to 1053 white and 1113 negro children, and the National Intelligence Test to 5834 white and 1112 negro children. The percentage of negro children who excelled the median of the white children at each age was as follows:

	Age								
	8	9	10	11	12	13	14	15	16
Myers Test	40	23	20	24	22	22	12	26	12
National Intelligence Test		10	28	31	17	17	21	20	15

From this comparison it appears that the language test does not place the negro children at a disadvantage. They stand low in both the language and the nonlanguage tests. In this respect they differ from some of the foreign language groups.

[1] Dagne Sunne, "Comparison of White and Negro Children in Verbal and Non-Verbal Tests," *School and Society*, XIX (April 19, 1924), 469–72.

A second racial group which is very distinct in the United States is the Indian. At least three studies of the standing of Indians in intelligence tests have been made, and they show uniformly that Indians make even lower scores than do negroes. Furthermore, in two of the experiments the pure blood Indians were found to stand lowest, while those of mixed blood stood higher in proportion to the amount of white blood in their veins. The following table from Hunter's [1] study is representative. The Indians were stu-

TABLE XXXVIII. SCORES ON THE OTIS TEST OF INDIANS OF DIFFERENT DEGREES OF MIXTURE OF BLOOD

PERCENTILE	¼ BLOOD	½ BLOOD	¾ BLOOD	PURE BLOOD
25	77.25	68.0	56.31	35.8
50	109.3	91.47	77.75	67.46
75	127.9	117.9	108.3	94.35
Number of cases...	112	192	142	265

dents at Haskell Institute and were therefore probably a somewhat select group. They represented 65 tribes and 14 tribal mixtures. The comparison of their scores with that made by whites may be expressed in the statement that 85 per cent tested below age. The correlation between the score and the amount of white blood, including pure whites, was found to be .51, when age and schooling were made constant.

Other racial comparisons have been made by testing the children of various immigrant groups. If the comparison is

[1] W. S. Hunter and Eloise Sommermeier, "The Relation of the Degree of Indian Blood to Score on the Otis Intelligence Test," *Journal of Comparative Psychology*, II (1922), 257–77.

made only between children of foreign born parents and American children of American parents the effect of limited familiarity with the English language upon the children's achievement in the tests must be considered; but this difficulty is not present when we compare children of different national origins with each other. When certain national or racial groups fall consistently low and others stand consistently well up, and when they are similar in respect to possible language handicap and unfavorable social environment, there appears to be good evidence that an inherent racial difference exists.

A typical study is the one by Murdoch.[1] She compared boys of four groups in the Pressey Group Scale of Intelligence. The results are shown in Table XXXIX.

TABLE XXXIX. MEDIAN SCORES OF CHILDREN OF FOUR RACIAL OR NATIONAL GROUPS

RACE OF GROUP	NO. CASES	AGE						
		9	10	11	12	13	14	15
Jews.....	500	109.0	109.3	118.4	128.5	125.5	124.3	126.5
Italians....	500	73.5	84.3	94.8	105.5	109.5	109.5	113.5
Americans	230	108.5	108.5	118.0	127.0	131.0	128.5	120.0
Negroes....	500	106.5	112.5	106.0	112.5	115.5	108.0	106.3

The Italian group is seen to stand lower than any of the other groups. Italian children and Polish children are found uniformly to stand low in studies of this sort. In this study but 15 per cent of the Italian children equaled the median score of the Jewish and American children.

The conviction that such differences as these are not

[1] Katherine Murdoch, "A Study of Race Differences in New York City," *School and Society*, XI (January 31, 1920), 147–50.

wholly due to language or environment is strengthened by the fact that Chinese children in the United States stand fairly high in tests. Pyle [1] gives the following comparison, showing the per cent which the Chinese children's average score is of the American children's average score.

	Boys	Girls
Rote memory	117.0	108.3
Logical memory	87.3	94.7
Substitution	88.6	77.9
Analogies	36.0	26.8
Spot pattern	90.4	

Pyle writes that the mentality of the Chinese children is much nearer the norm for American white city children than is that of negro children or of rural whites. He believes that if allowance were made for language difference the Chinese children would equal that of American whites.

Later studies have confirmed these findings and none, so far as the author is aware, has contradicted them. For example, Goodenough [2] classified the scores of 2457 children from different parts of the country on her drawing scale, which has been found to be a measure of general ability. In this tabulation the following groups made average or higher than average scores: Americans, Jewish, Chinese, Japanese, Germans, English and Scotch, Danish, Swedish, and Norwegian. The following groups, on the other hand, made scores below the average: Armenians, Italians, Spanish, Mexican, Negroes, Indians, Portuguese, French and Swiss, and Assyrian, Slovenian, and Serbian. Since this test would not be expected to be very greatly influenced by

[1] W. H. Pyle, "A Study of the Mental and Physical Characteristics of the Chinese," *School and Society*, VIII (August 31, 1918), 264–69.

[2] Florence L. Goodenough, "Racial Differences in the Intelligence of School Children," *Journal of Experimental Psychology*, IX (October, 1926), 388–97.

environment and since these differences are consistent with those which have been found on other tests the evidence suggests that there are native racial differences. We must be on our guard, however, against assuming that the amount of these differences is as great as is indicated by the test scores.

5. Differences between groups with various amounts of schooling

We have had illustrations of differences between persons working at different occupations, living in different places or belonging to different races. Our additional comparison will lead us directly to the interpretation of the facts which have been reviewed. It has been pointed out repeatedly that there is a correlation between the amount of schooling an individual has had and his standing in intelligence tests. The army tests give us the most extensive data on this point as on many others. In Table XL[1] are compiled the median scores of groups of men who are shown by their

TABLE XL. THE AVERAGE SCORES IN ARMY ALPHA MADE BY MEN WITH DIFFERENT AMOUNTS OF SCHOOLING

GROUP	SCHOOL GRADE COMPLETED					
	0–4	5–8	High School	College	Beyond College	TOTAL SCORE
White officers	112.5	107.0	131.1	143.2	143.5	139.2
White draft native	22.0	51.1	92.1	117.8	145.9	58.9
White draft foreign	21.4	47.2	72.4	91.9	92.5	46.7
Colored draft, North	17.0	37.2	71.2	90.5		38.6
Colored draft, South	7.2	16.3	45.7	63.8		12.4

[1] Robert M. Yerkes (Editor), *Psychological Examining in the United States Army*, assembled from tables on pp. 706, 767, 768, and 770.

records to have completed respectively four school grades, eight grades, the high school, and college. This comparison is made for five groups of men.

The outstanding fact which is revealed in this table is that, with one exception in the case of the officers, the men who have had more schooling make the higher scores. This appears when we compare the averages in the various horizontal rows, looking from left to right. The table also gives us a comparison between the different groups of men, which can be made by running up and down the vertical columns; but our primary concern is with the first comparison.

A more illuminating comparison than the mere contrast in the test scores of children with different amounts of schooling is revealed when this difference is traced through successive ages. This has been done in several cases and in each comparison the deficiency on the mental test becomes progressively greater as children grow older. For example, a study was made by Hugh Gordon of canal-boat children in England. These children live with their parents on the canal boats and therefore have little formal schooling. Gordon found that their average I.Q. dropped from about 90 at six years of age to about 60 at twelve years of age. A similar drop has been found in the case of isolated mountaineer children of the South. Sherman found, for example, in tabulating the scores of children from six to twelve years of age that the youngest group had an average I.Q. of 83, the middle group of 70, and the oldest group of 50. Wheeler,[1] in a study of East Tennessee mountain children, tabulated the scores of 564 children by age. He found that there was a decrease from an average I.Q. of 92.5 at age nine to 72.5 at age fifteen on the Illinois Examination and from 95.3 to

[1] L. R. Wheeler, "The Intelligence of East Tennessee Mountain Children," *Journal of Educational Psychology*, XXIII (May, 1932), 351–70.

72.5 on the Dearborn Intelligence Test. This indicates that as the difference in schooling becomes progressively greater the difference in the scores on the tests also becomes progressively greater.

6. Interpretation of the various group differences

In our effort to get an explanation of the various group differences which have been shown, and by so doing to gain light on the fundamental meaning of intelligence tests, let us begin with the last comparison. The two contrasted views of the meaning of the tests are well represented in the explanations which are offered of this simple fact. The first view is the apparently simpler one that persons with greater amounts of schooling make higher scores because their schooling raises their intelligence, that is, increases the ability which is measured by the tests. The mere mathematical fact of correlation, of course, does not indicate which is cause and which is effect, but the suggestion which comes most naturally to mind is that schooling is the cause and intelligence is the effect. This is the explanation which many hold to be the true one.

Many psychologists, on the other hand, believe that the explanation which is apparently the simpler is proven by other facts not to be the true one. We know, for example, that pupils drop out of school in part because of a limitation in their ability, which makes it difficult, if not impossible, for them to go on. One psychologist, Pillsbury, regards this selective elimination as of such importance that he considers the most important function of the school to be to pick out the intelligent individuals and put them in positions of leadership, rather than to teach them.[1] The view that the chief factor in the correlation between intelligence and

[1] W. B. Pillsbury, "Selection — An Unnoticed Function of Education," *Scientific Monthly*, XII (January, 1921), 62–74.

amount of schooling is native capacity is further supported by the other evidences that intelligence test scores are dependent chiefly on native capacity. We may proceed to consider some of the other evidences which lie in the data before us.

Consider first the exception in the above table which appears in the case of the white officers. Those who have had only four years or less schooling stand higher than those who have completed five to eight grades. This fact is puzzling on the hypothesis that it is education which is solely responsible for high scores. We seem here to have a few very intelligent individuals who, in spite of extremely limited education, are chosen as officers and make relatively high scores on the test. On further examination, however, there is evidence in these data that the amount of education affects the score. This group of men with very little schooling must have been at least as gifted by nature as the aver-

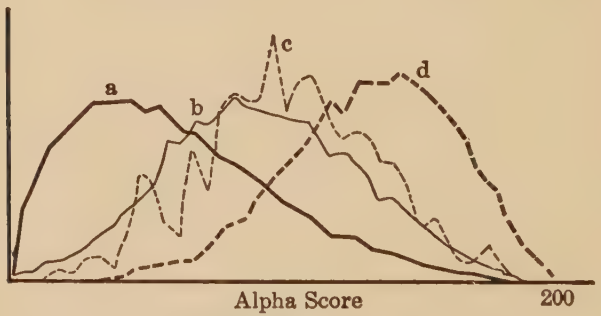


FIG. 18. DISTRIBUTION OF ALPHA SCORES OF OFFICERS AND ENLISTED MEN WITH MORE THAN EIGHTH-GRADE OR LESS THAN EIGHTH-GRADE SCHOOLING

(From *Army Report*, pp. 765, 779)

age of the officers. Otherwise they would not have been made officers in spite of their meager education. Their low score in comparison with the high-school and college students, then, must be due to their lack of education.

The same reasoning applies to the "crucial test" of the army tests which is described by Brigham.[1] Brigham cites the *Army Report* (pp. 778, 779), which shows that 660 officers who had not gone beyond the eighth grade in school made slightly higher scores on the average than 13,943 native born recruits (not officers), all of whom had gone beyond the eighth grade. The officers, in spite of their lack of education, made slightly higher scores than enlisted men with more education.

This fact does adequately prove that native capacity is a large factor in the intelligence test; but it shows with equal clearness that education is also a factor. The officers in question must have been by nature even more intelligent than the average officer. Now officers in general made so much higher scores than enlisted men that while 83 per cent of the officers received A and B grades, but 18.8 per cent of the enlisted men made these grades.[2] The median score of the officers with less than eighth grade education was 107.3, while the median score of the enlisted men with more than eighth-grade education was 97.4, a difference of only 9.9 points. When their education is equal, the superiority of the officers is very much greater. The group of officers who completed high school (see Table XL) had an average score of 131.1 while the corresponding group of enlisted men had an average score of 92.1, a difference of 39 points. Deficiency in education reduced the first group from a superiority of 39 points to a superiority of only 9.9 points.

The "crucial" comparison of the privates and officers in the army is extended in Fig. 18 so as to include four groups;

[1] Brigham, *op. cit.*, pp. 63 ff. [2] Yoakum and Yerkes, *op. cit.*, p. 27.

all enlisted men, enlisted men of more than eighth-grade education, all officers, and officers of less than eighth-grade education. We can now make three comparisons. First, we can compare enlisted men in general (Curve *a*) with officers in general (Curve *d*). This comparison yields us little of value, since the two groups differ both in schooling and intelligence, and it is impossible to tell how much the difference in scores is due to the one and how much to the other. In the second place we can compare men of the same rank but of different amounts of schooling. It is evident that the enlisted men with superior schooling (Curve *b*) make much higher scores than all enlisted men, whose schooling is much less (Curve *a*). Similarly the officers with inferior schooling (Curve *c*) make much lower scores than officers in general, whose schooling is much more (Curve *d*). Obviously schooling makes a large difference in the scores. Finally, we can compare officers with little schooling (Curve *c*) with enlisted men of much schooling. The higher scores of the officers indicate that native capacity makes a large difference in the scores. These two latter comparisons indicate unequivocally that the Army Alpha scores were affected, perhaps about equally, by differences in schooling and by differences in intelligence.

It is evident from a review of these group comparisons that in almost every instance it is possible to make an opposite interpretation of the facts. This is because it is possible in almost every instance that there are differences in both heredity and environment between the groups which are compared. If differences in environment exist between the groups, it is possible also to assume that there are differences in hereditary constitution in the same direction as the differences in environment. Since the two factors may vary and may vary in the same direction, it is a matter of speculation as to which one is responsible for the differ-

ences in ability as indicated by the tests. In order to secure more conclusive evidence, therefore, of the relative effect of these two sets of factors in producing individual differences it is necessary to find instances or to make instances in which one factor varies while the other factor remains constant. The method of partial correlation would enable us to estimate the effect of one independently of the other if we had independent measures of the two factors. We may, however, be able to find groups which are sufficiently clearly contrasted in respect to the two factors so that a more crucial comparison can be made. There are three types of cases in which the evidence appears to be more conclusive than in the majority of the group comparisons which have been described. These are, respectively, studies of the effect of special variation in education, studies of the intelligence of foster children with the usual relation between children and parents absent, and studies of twins.

The most prominent example of the study of the effect of special type of education is the series of investigations made by Wellman which has already been cited. The general finding of these studies was that the children who attended the pre-school at the University of Iowa made large gains in intelligence and that these gains persisted until the college period. An analysis of these gains in addition to the general fact suggests strongly that they are to be attributed to the pre-school education of these children. For example, (1) the gains took place during the academic year and not in the summer; (2) the amount of gain was associated with the length of attendance within the year; (3) pupils who transferred to other schools maintained their gain but did not increase it; (4) children who remained in the University Elementary School continued to gain; (5) the gain was associated with the length of

interval between the first and last tests; (6) non-pre-school children did not show the gain shown by the pre-school children nor did they show the contrast between the gain in the school year and the gain in the summer; (7) bright children tended to lose and duller children to gain contrary to Cattell's finding. It must be acknowledged that other studies of the intelligence of pre-school children in comparison to those who do not attend pre-school have failed to show a similar superiority. This, however, may be due to the type of training given in the particular pre-school attended by the children. At any rate, it is difficult to explain away positive results such as those found by Wellman.

Several studies have been made of the intelligence of foster children. Companion studies made at Stanford University and at the University of Chicago were reported in the Twenty-seventh Yearbook of the National Society for the Study of Education. Both of these studies showed that there is a correlation between the intelligence of foster children and the rating of the foster homes in which they are situated. This correlation may conceivably be due to selection, at least in part, but an analysis of the situations in both studies appeared to indicate that selection was not the sole cause. In the University of Chicago study certain other comparisons were made which seemed to indicate that the environment of foster homes was in part responsible for the intelligence of the children. For example, the average intelligence of the entire group of 401 children was practically 100. It is certain that the intelligence of the parents was considerably below 100. This suggested, at least, a beneficial influence of the homes. Again, a group of 75 children who were tested before and after adoption gained in I.Q. and those who were in better homes gained more than those in poorer homes. If only brothers

and sisters were compared and if they were put into two groups according to the homes in which they were placed, it was found that those in better homes had higher I.Q.s than their brothers and sisters in poorer homes. These and other facts suggest that the home environment had an influence on the intelligence of the children.

Still more striking evidence is found by Skeels [1] in his study of foster children at the University of Iowa. At the latest report Skeels had tested 147 foster children who had been adopted in early infancy. The average I.Q. of these children was 115.4. The average of those below twenty-four months of age was 119.0, and those above twenty-four months, 107.9, suggesting that the test may have had something to do with the score. However, the mean I.Q. of 78 of the true mothers was found to be only 87.0, and there was no correlation found between the I.Q.s of the mothers and of the children. It is difficult to avoid the conclusion that the influence of the foster home was responsible in large measure for the I.Q.s of these children.

Perhaps the most crucial comparison made up to the present is that between twins. Representative findings will be cited from the study on twins by Newman, Holzinger, and the writer.[2] The study consists of two parts, in each of which a major comparison was made. In the first part, fifty pairs of fraternal and fifty pairs of identical twins who were reared together were tested and studied. In this part of the study the environment was about the same

[1] Harold M. Skeels, "Mental Development of Children in Foster Homes," *Pedagogical Seminary and Journal of Genetic Psychology*, XLIX (September, 1936), 91–106; *Journal of Consulting Psychology*, II (March–April, 1938), 33–43.

[2] Horatio H. Newman, Frank N. Freeman, and Karl J. Holzinger, *Twins: A Study of Heredity and Environment.* Chicago: University of Chicago Press, 1937.

between the members of each pair of twins but the contrast exists between the group in which there is identity of heredity, namely, the identical twins, and the group in which the resemblance is only half as much, that is, the fraternal twins. In the case of these twins, then, we have an opportunity to study groups which are similar in the environmental influence but are contrasted in the hereditary or genetic factor. In the second part of the study, nineteen pairs of identical twins who were separated in infancy and reared apart were studied. In this case, the heredity of the members of each pair was the same but the environment differed. In this case, then, the comparison is based on differences in environment with similarity in heredity.

Space will not be taken for a detailed presentation of the findings of the study of twins. They may be summarized in a table showing the correlations between twins of the three groups. We may compare, first, the correlations of

TABLE XLI. CORRELATIONS FOR THREE GROUPS OF TWINS

Trait	Identical	Fraternal	Separated
Standing height	.981	.934	.969
Sitting height	.965	.901	.960
Weight	.973	.900	.886
Head length	.910	.691	.917
Head width	.908	.654	.880
Binet mental age	.922	.831	.637
Binet I.Q.	.910	.640	.670
Otis I.Q.	.922	.621	.727
Stanford Achievement	.955	.883	.507
Woodworth-Mathews	.562	.371	.583

identical and of fraternal twins. In all cases the correlations of fraternal twins are lower, showing that they resemble each other less than do identical twins. The difference is greater in some traits than in others as, for example, head length and width, Binet mental age and I.Q., Otis I.Q. The

difference is not so great in weight and in Stanford Achievement. The lower resemblance in general between fraternal twins indicates that their lesser communality in genetic constitution is responsible for a lower degree of similarity because both types of twins are reared in the same homes and therefore under similar environmental conditions. If now we compare the correlations of the separated twins with the other two groups, we find that they are more like the identical twins who are reared together in some traits, and more like the fraternal twins in others. For example, in the physical measures of standing height, sitting height, head length and head width their correlations resemble those of identical twins reared together. In the physical trait of weight, on the other hand, and in the mental traits of Binet mental age, Binet I.Q., Otis I.Q., and Stanford Achievement the resemblance is more like that of fraternal twins than of identical twins. It appears, therefore, that a diversity in environment affects these traits even when the comparison is between individuals who have exactly the same heredity. This conclusion is strengthened when we find that the difference in intelligence of these separated identical twins corresponds very closely with the kind and amount of difference in schooling between them. In other words, differences in environment do produce differences in ability.

7. Summary

This detailed examination of the scientific evidence which is at hand indicates the correctness of the moderate view as contrasted with either extreme. As was pointed out in the first chapter, one may regard intelligence tests as an entirely new and perfect instrument for detecting native capacity. At the other extreme he may discount them and regard them as merely somewhat improved instruments for

measuring the results of teaching. The consideration of the historical development of tests, in common with an analysis of their results, shows that neither of these views is correct. Intelligence tests have made a marked advance toward the measurement of native capacity, but their scores are still influenced to a considerable degree by the effects of training, and in their interpretation this influence must always be taken into account.

Chapter XVI

THE NATURE OF ABILITY

In the foregoing description of the historical development of mental tests, of the various types of tests, and their uses and interpretation, we have frequently encountered the question of the nature of the ability which is measured by these tests. This question is sometimes ignored and the position taken that it is possible to use mental tests for practical purposes without determining what it is that they measure. The question of what is measured by the tests does, however, have a bearing upon their design, use, and interpretation and it is necessary, therefore, at the conclusion of this volume to review the main concepts of the nature of ability.

In the early history of the testing movement the problem of the nature of abilities was not acutely felt. The earlier tests were designed to measure rather narrow abilities which were defined in common-sense terms or in terms of ordinary psychological concepts. It was assumed that these abilities could be measured by tests which appeared from common sense to call for the exercise of the ability. The question was not raised whether the ability was generalized, that is, whether all the activities that appeared to represent the ability were equally good representatives of it. To put it in present-day terms, it was not asked whether there was a correlation between the various acts which would commonly be classed under the same ability. Thus, tests of manual dexterity were given without asking whether there

is such a thing as manual dexterity in general or whether what we call manual dexterity is made up of a host of particular abilities. Because the early tests dealt with rather simple abilities the existence and nature of these abilities was assumed rather than inquired into.

It was when the question of testing higher types of abilities, those which would ordinarily be called more intellectual, arose that theories began to be formulated concerning the nature of this ability. Two lines of development then began. The first is represented by the work of Binet. Binet appeared to think of intelligence as a kind of composite of a considerable number of types of performance or of the ability to carry on a number of types of performance. At the same time he seemed to regard the ability to carry on these various types of performance as an indication of an underlying characteristic which was not to be identified with any one of them. It was expressed or represented in various particular forms of activity and could best be measured by giving opportunity to carry on a variety of these particular forms. The successors of Binet have likewise refrained from attempting to formulate any exact definition of intelligence. They have been content with a general description of the sorts of things that intelligence enables one to do.

The second line of development was initiated by Spearman and has been carried forward by him until the present. This development rests upon statistical analysis which has recently been pursued vigorously under the name of factor analysis. The aim of those who carry on factor analysis is to distinguish sharply between various abilities, to define them, and ultimately to develop tests to measure them. The factor analysts have criticized the psychologists of the Binet school on the ground that their procedure is vague and undefined, and empirical rather than psychological.

Along with the development of these attempts to measure general ability in the form of general intelligence or something similar to it has gone the development of tests for various other types of abilities. In the chapter on "Tests for the Analysis of Mental Capacity" we traced some of these tests, such as the tests of aptitude. The definition of aptitudes, as we have seen, has been largely based upon the demands of practical activities. This procedure, therefore, has been empirical in the same sense as the attempt to measure intelligence by such scales as the Binet scale. When attempts have been made to measure other abilities, such as memory, the concept of these abilities has usually been borrowed from common sense or from ordinary psychology. The tests to measure them have usually been selected by similar exercise of common sense.

1. Concepts and theories of abilities

We may now trace briefly various concepts of abilities and attempt to evaluate them in so far as our present evidence allows.

In the consideration of these conceptions of ability we may keep in mind three points. In the first place, we may inquire about the descriptive character of the abilities: What are they like? In what forms of activities are they represented? How may we recognize them? In the second place, we may ask, What is the organic basis of the abilities? or at least, What theory is held concerning the organic basis? In some cases such theories exist and in other cases they do not. Finally, we may inquire whether the theory involves the belief that the abilities are native or are acquired, or whether it says nothing on this question.

2. The faculty theory

The first definite theory concerning the organization of abilities was the theory of faculties. According to it, ability in general is composed of a large number of elements or components. Each one of these exists in varying degree in each individual. They may or may not be correlated with each other, but they were commonly thought of as being relatively independent of each other. Some of these abilities are rather broad and general, such as memory and imagination, some are more specific, such as ability in language or in number, and some are not abilities at all but personality characteristics or traits, such as caution or impulsiveness. According to the faculty theory the various abilities are localized in particular areas of the brain. In the particular form of the theory called phrenology it was held that development of these brain areas could be detected by feeling the bumps on the skull. It is not quite clear whether the faculty theory involves the belief that the faculties and the local areas in the brain were determined by inheritance or whether they were acquired. The theory of formal discipline, which was sometimes connected with the faculty theory, however, assumed that the various faculties and therefore the areas of the brain could be developed by exercise.

The faculty theory fell into disrepute on account of attacks from two directions. It was opposed in the first place by the particularist or atomistic view of learning and abilities promulgated by Thorndike. According to this view abilities are to be thought of as exceedingly narrow and specific elements of behavior. The ability is defined not merely in terms of a psychological entity or element but in terms also of the particular situation to which the response is made. In other words, every stimulus response unit would constitute an ability and the element of the situ-

ation to which the response is made would be a part of the definition of the ability. Ability, in general, would be made up of all of the particular abilities which the individual possessed. If there is such a thing as general ability, it is merely the sum total of all of the particular abilities.

According to this view of specific abilities the basis of these abilities in the brain is the element of the nervous system which is required to perform the activity which represents the ability. An ability, then, is based on a group of small neural elements or a group of neurones. Broader or more general abilities would merely be based on larger groups of neural elements.

Whether abilities as defined in this sense are inherited or acquired depends upon the predilection of the person holding this theory. Abilities might be regarded and are regarded by various proponents of the theory as inherited, and by others as acquired. Behaviorists, for example, hold some such theory as this and they regard abilities as acquired. Thomson, on the other hand, considers that they may be inherited according to the Mendelian Law.

3. The two-factor theory

The first of the theories to grow out of the modern studies in correlation is the two-factor theory of Spearman. This is distinct from previous theories in defining abilities not in terms of activities which are represented in everyday life or in the activities which correspond to given tests, but in terms of factors which underlie these activities. It may be possible, according to Spearman, to find activities which correspond closely to these factors, and to construct tests which will demand and measure the performance of these activities. One discovers these factors and the means of measuring them, however, only through the studies of correlation. In his first formulation Spearman described two

factors, and his theory was therefore called the two-factor theory. The first of these is called "*g*." It is general in that it enters into every kind of activity. The second factor he called "*s*." It is the special factor or factors in an activity which combines with *g* to constitute the total activity. Spearman showed that one could account for the correlation between tests or the interrelationships between abilities by assuming the existence of *g* and the various *s*'s in all cases in which the tetrad equations are zero, that is, in cases in which equal proportionality exists between the correlation coefficients of a set of tests. He very soon came upon cases, however, in which this is not true. In these cases groups of tests correlate more highly with each other and can be accounted for by their possession of *g*. For these he posited the existence of additional factors, called group factors. These are less general than *g* in that they do not run through all the tests or abilities, but they are more general than the *s*'s in that they appear in a group of tests or a group of abilities.

Spearman goes on to describe *g* on the basis of his analysis of the tests which appear to be more highly saturated with *g*. *G*, according to his description, consists of two types of relational thinking which he calls the eduction of relations and the eduction of correlates. The eduction of relations is illustrated by the apprehension of the relations existing between two objects or parts of an object. The eduction of correlates is carried on when one has in mind an object and a relation and thinks of another object which has that relation to the first one. Group factors are illustrated by number ability, verbal ability, mechanical ability, and mental speed. Spearman, then, has finally three kinds of abilities, *g*, which is perfectly general and pervades all activities, *s*, which is highly special and belongs to only one activity, and the group factors, which are intermediate between the two.

Spearman has a special theory of the organic basis of abilities. *G*, he thinks, is based upon the amount of neural energy which is at the disposal of the individual for carrying on intellectual operations. The *s*'s are represented in the structure of particular parts of the nervous system. These he calls the "engines" in contrast to *g* which represents the energy. So far as the writer is aware, Spearman has not advanced a definite explanation of the organic basis of the group factors, but it would probably consist in structures in the brain of wider scope than those that underlie the *s*'s.

Spearman refrains from committing himself as to whether abilities are dependent on inheritance or on training. He appears, however, to lean to the view that they are, at least in large part, due to inheritance.

4. Primary abilities

Until recently the field was largely occupied by the two-factor theory of Spearman amplified by Holzinger into the bi-factor theory by including group factors. Some factor analysts, including Kelley and Thurstone, believe that the correlations between tests can best be accounted for by assuming the existence of a limited number of abilities which Thurstone calls "primary abilities." These abilities, as was seen in the discussion of factor analysis, are similar in their descriptive nature to the faculties of the older psychology. This similarity is evident from the list given by Thurstone: number facility; word fluency; visualizing; memory of words, names, and numbers; perceptual speed; induction; and verbal reasoning. The newer lists of abilities differ, of course, from the old faculties in that they are based upon an attempt to account for the correlations which are found between tests instead of being based upon ordinary observation or common sense.

Modern factor analysts other than Spearman have not, so

far as the writer is aware, attempted to explain the organic basis of the abilities which they posit. They doubtless assume that there is an organic basis but have not attempted to define it. They have not, furthermore, made a definite pronouncement as to whether the abilities in question are native or acquired. They appear to have leaned, however, to the doctrine that they are at least largely if not entirely native. This is suggested, for example, by the fact that the analysis of primary abilities has sometimes been used in the effort to give vocational guidance.

5. Criticism and evaluation of the doctrines

Some psychologists are disposed to be critical of the whole doctrine of abilities which has been derived from factor analysis. They object that these doctrines set up entities in the brain or in the mind, and that such entities are not in accord with our knowledge of the way human activity is carried on. The factor theories appear to assume that abilities are uniform within themselves and may be distinct or independent from other abilities.[1] Behavior, on the other hand, seems to shade by small degrees from one activity to another; and different forms of activity appear to have a good deal in common with each other. It seems to some, therefore, a psychologically false interpretation to set up independent and separate entities which are different from each other but the same in themselves wherever they appear. This conclusion would seem to be necessary of the factors which are assumed to be responsible for correlations. Another way of expressing the criticism is to say that the factors seem to be mathematical factors rather than psychological realities.

It is, of course, not a valid argument against the theories

[1] This assumption is not necessary nor is it always made. The subsequent discussion is based on the assumption of independence.

based on factor analysis to point out that tests which seem, on the surface, to measure a given ability do not correlate closely with each other or that tests which do not, on the surface, seem to measure a given ability do correlate with each other. The factors may be hidden, and they may be associated in a given test with others which influence the score. Only if the test is relatively pure, that is, saturated with one factor and free from any other factor, would we expect it to correlate highly with other tests of the same factor and not at all with tests of other factors. It must be admitted that the possibility of tests of this sort is largely hypothetical. Although, on the one hand, the failure of our present tests to correlate as we should expect if ability is composed of clear-cut factors is not a conclusive argument against the theory of factors; on the other hand, the existence of such factors has not yet been demonstrated concretely by the development of tests which represent them in pure form.

If it is argued that the theory of factors does not imply that such tests can be devised and that we should not demand them as evidence, it may be pointed out that the factor analysts themselves have almost universally named and described their hypothetical factors in terms of the usual categories of descriptive psychology and even of common sense. This is a surprising result. We might have expected that factor analysis would reveal factors that would be inaccessible to ordinary observation, something like the air waves, light waves, atoms, molecules, electrons, or cosmic rays of physics. But we do not seem to reach new concepts from our exploration of abilities by factor analysis. It is a fair question to ask whether the similarity of the concepts we do come out with to those of common sense is not due to the fact that they are what we went in with. If there is any ground for this interpretation of the procedure

and conclusions of factor analysis, and if our psychological conceptions may in some measure control the outcome, we are justified in seeking to refine our conceptions of the nature of ability on the ground of general psychological and biological experience. The facts of correlation constitute valuable evidence with which to construct a theory, but not the sole evidence.

A fundamental question to be considered in building our conception of abilities is the question whether they are inherent or acquired. If the conclusions of the preceding chapter are correct, individual differences in ability are partly inherent and partly acquired. If this is the case, the problem of the constitution of abilities is indeterminate. The conditions of the environment vary so greatly that we should not expect to find complete uniformities from person to person in the co-existence of degrees of ability. The relations would vary with relative degrees of emphasis in training and experience. It seems probable that such is the case. There might still be underlying uniformities, but they would be overlaid with modifications by the environment. Whatever trait groupings might be found could never be determined exactly.

If we admit, as we seem compelled to do, that experience modifies abilities and renders the measurement of native ability indeterminate, may we perhaps go further and conclude that experience actually creates the abilities which are manifested by the correlations? To say this does not mean that individual differences in native ability may not exist, but rather that the groupings of abilities or traits which we call factors may be created by the habits which are formed in reacting to given environmental situations or demands. Language, for example, has been identified by factor analysis as a component of ability. Is the capacity for language a native unitary ability, or is the union of the

elementary processes which make up the whole complex process of language — distinguishing sounds, uttering sounds, grasping meanings, connecting sounds and pronunciation with meanings, etc. — a product of the social environment? It is conceivable that these elements do not belong together by nature and are not inherited by the individual, but that they are rather put together in the experience of each individual because language is a characteristic form of behavior of the society in which he lives, and because the acquisition of this form of behavior requires him to develop the system of activities which constitutes it. Language, in such case, would be a social category and not an individual one, or perhaps it would be better to say that it becomes an individual category only as the individual takes it on.

The theory of factors or abilities is not yet sufficiently explored to yield a certain answer to such a question, and this view is offered only as a tentative hypothesis. It is suggested because the alternative hypothesis encounters difficulties. It resembles too closely the faculty theory, which psychologists have long rejected. It is hard to see that it makes any difference whether memory, language, mathematical facility, spatial imagination, etc., are called faculties or primary abilities. Certainly the phrenological conception that faculties depend on the structure and development of localized areas of the brain is not admissible. It is hard to conceive a form of brain structure that would explain native individual differences in abilities like the ones named. They could more easily be explained as organized habits or modes of behavior.

Some individual differences could readily be explained as due to inherent characteristics of the brain, nervous system as a whole, or other parts of the body. For example, sensory acuity may plausibly be said to depend on the

structure of the sense organs. Motor ability may possibly depend on the structure of the muscles or some general quality of the brain and nerves which facilitates conduction and coördination. This latter conception is admittedly vague, but it is not contrary to known facts, as is the conception of localization of complex functions. General intellectual ability may, as Thorndike suggests, be based on the number of neurones of the brain which determines the number of possible associations, or on the total store of energy available for brain action, as suggested by Spearman. Either of these suggestions fits better the notion of intelligence represented in a variety of intellectual operations than does Spearman's conception of a general factor consisting only of two processes, eduction of relations and eduction of correlates. Intellectual ability is doubtless best exhibited in more complex rather than in simple operations, in novel rather than in reflex or habitual forms of adjustment, in abstraction and generalization rather than in manipulation of the particular and concrete. This conception fits well the character of the tests which have proved most satisfactory as constituents of measures of general intellectual ability.

General intellectual ability as so conceived may rest partly on the inherent quality or structure of the nervous system, particularly the cerebrum. The intellectual superiority of man over the animals is commonly believed to lie in the greater size and complexity of his cerebrum. Differences between individual human beings can most plausibly be attributed to differences in structure of their cerebrums. Size has been found to have but low correlation with intelligence, but extremely small brains are indicative of mental deficiency. Intelligence seems to be associated with the number and size of the arteries in the covering of the brain, the pia mater, which indicates that the efficiency of

functioning of the brain depends on the abundance of its blood supply. These may constitute the physical basis for what psychologists call general intelligence and what the man on the street calls brain power.

So far as native abilities are concerned, then, it is most plausible to posit the existence of differences in sensory acuity, possibly in motor dexterity, and probably in general intellectual ability, with emphasis on the so-called higher processes. Native differences in the primary abilities or group factors may also exist by native constitution, but they are a less plausible explanation because they do not, like sensory capacity, rest on known specific structures, or, like general intelligence, on some general quality or characteristic of the brain as a whole. They may, on the other hand, be readily explained as forms of behavior acquired in response to organized sets of environmental social stimuli.

The conception that ability is made up of a large number of narrow components readily explains the statistical facts of correlation, when the correlations are not too high, and agrees with theory of heredity as mediated by groups of genes. This conception, however, encounters a psychological difficulty in assuming that an ability is made up of a large number of disparate elements, or at least of differences in ability due to differences in the chance groupings of many elements. Studies of heredity in human beings are difficult and their interpretation uncertain, especially in mental traits. We can, therefore, rely little on the present evidence from heredity as an indication of the constitution of abilities. Future studies may give fuller information. If it should later be possible to determine how mental abilities are inherited, we could infer from the combinations of inherited traits the factors or elements of which they are composed. At present we can only make uncertain inferences from remote analogy with physical traits. Such inference is insecure.

The foregoing speculations concerning the nature of ability, like all speculations on this subject, are made on a meager foundation of fact. They constitute one more attempt, added to all that have gone before, to find the conception of abilities that will accord best with the statistical, psychological, and biological facts as they have been ascertained up to the present. The conclusions reached are, in brief, as follows: Certain simple abilities, mainly sensory, are relatively independent, largely inherent, and based on the structure of the sense organs. Motor ability may be also based largely on structure, but the structural basis is not so specific; motor ability is therefore not unified. General intellectual ability is a reality and there are two or three plausible explanations of its organic basis. It is not to be identified with any one particular thought process, but includes a number of modes of thinking, with emphasis on the more abstract and complex ones. The group factors or primary abilities are probably not native abilities but habits of thought formed by experience and training. All abilities, both special and general, are affected by training, but some appear to have an underlying organic basis.

Index

Abernethy, Ethel Mary, and relation of mental and physical growth, 312; *cited* 312
Ability, tests for general or special, 26; specialized tests of, 239
Ability groups, classification into, 376–384
Academic aptitude, tests for, 192–193
Accomplishment ratio, 27, 303
Accuracy, of score, 279
Achievement and intelligence, relation between, 303–308; correlation with mental tests, 353–362; mental tests as measures of factors in, 362–363; and community intelligence level, 370–373
Achievement quotient, 27; defined, 153; and accomplishment ratio, 303
Adaptation, Binet's test for, 53
Adaptation to age or grade, a criterion in choice of test, 157
Adjustment of pupil to his work, uses of tests in maintaining, 391–393
Administration of tests, technique of, 73; ease and simplicity, a criterion in the choice of a test, 159
Administrative use of mental tests, 373
Adult tests, listed, 167–168
Age, a factor affecting test scores, 66
Age norms, 309–312
Age progress curves, 291
Age scales, 85–106; Binet's early work, 85; scores on measure intelligence, 87; Binet-Simon 1908 scale, 88–90; 1911 Binet-Simon revision, 90–92; Goddard's revision, 92–93; Kuhlman's revision, 93–94; Stanford revisions, 94–106; and point scales, 108–112
Aims, of mental tests, 19
Alexander, Herbert B., and geographical distribution of Army Alpha scores, 409; *cited* 409
Allport, Gordon W., and Vernon, Philip E., and tests of attitudes, 227–228; *cited* 235
Allport, Gordon W., and Floyd H., and social reaction tests, 223; *cited* 234
Allport, Floyd H., 223; *cited* 234
Almack, John C., *cited* 236
Alternative tests, the, 272, 275
American Psychological Association, and early development of tests, 39; and Committee of 1906, 59; and army mental tests, 113–114
Analysis of mental capacity, tests for, 169–204; test groups, 170; Healy-Fernald test group, 171; other test groups, 178; aptitude tests, 182; mechanical aptitude tests, 183; musical aptitude, 187; art aptitude, 189; clerical aptitude, 190; academic aptitude, 192; special abilities, 194; profile tests, 198
Anderson, Helen M., *cited* 411
Anderson, J. E., *cited* 136
Anderson, Rose G., *cited* 165, 166, 167, 168
Andrews, Dorothy M., and Paterson, Donald G., and clerical aptitude tests, 191; *cited* 191
Appeal to the child, a criterion in choice of test, 158
Application of the correlation method, 59–84; Spearman's criticism of statistical procedure, 62–71; correlation studies of single tests, 71–78
Aptitude, defined, 182
Aptitude tests, nature of, 182; mechanical, 183–187; musical,

187–189; art, 189–190; clerical, 190–192; academic, 192–193; special abilities, 194–198; profile scales and, 198–204
Army, U. S., mental tests, development of, 113–114; Army Scale Alpha, 114–129; Army Scale Beta, 130–135; performance scale, 135–136; uses of, in army, 136–140; intercorrelation of, 251–253
Army Performance Scale Examination, described, 135–136; *cited* 166
Army Scale Alpha, in World War, 17; described, 114–129; and letter-rating, 125–127; and mental ages of army men, 128; development of, 128; revised form, 151; Scrambled Alpha, 152; speed *vs.* power in, 262–265; and race norms, 318; scores of occupational groups on, 405; scores of geographical groups on, 409; scores of immigrant groups on, 412; scores of racial groups on, 414; scores of men with different schooling, 419; scores of various ranks, 422; *cited* 167
Army Scale Beta, and Army Alpha, 114; described, 130–134; results of, 135
Art, tests for aptitude in, 189–190
Arthur, Grace, and Arthur Performance Scale, 147; *cited* 164
Association processes, tests dealing with, 55
Association tests, Burt's, 72
Attention, Binet's test of, 52, 53
Attenuation, correction for, 70
Attitudes, tests of, 226; list, 235
Audiometer, Seashore, 71

Bagley, W. C., and motor index, 44–46; and intelligence of racial groups, 414; *cited* 414
Baker, Harry J., *cited* 234, 235
Baker, Harry J., and Crockett, Alexander C., and mechanical aptitude tests, 185
Baker, Harry J., Kaufman, H. J., Engel, Anna M., and Detroit Kindergarten Test, 150; *cited* 165, 166, 167
Baker, Harry J., and Leland, Bernice, and profile scales, 201–203; *cited* 201, 202
Barry, Herbert, Jr., *cited* 234
Bayley, Nancy, and California First-Year Mental Scale, 149; *cited* 149, 164
Behavior, tests of, 215–220; list of, 234; relation of intelligence to, 363–369
Bell, Hugh M., and tests of neurotic tendencies, 225; *cited* 235
Bernreuter, Robert G., and Personality Inventory, 232; *cited* 236
Berry, Charles S., and horizontal classification of pupils, 378; *cited* 378
Bi-factor method, of Holzinger, 82, 437
Binet, Alfred, 2; early experiments, 51–55; and age scales, 85–92; 1905 scale, 85–88; death of, 91; definition of intelligence, 248; and nature of intelligence, 432; *cited* 53, 91.
Binet, A., and Simon, T., scale of 1908, 88–90; 1911 revision, 90–92; other revisions: Goddard's 1911 revision, 92; first Stanford revision, 94–98; the revised Stanford-Binet scales, 103–106; *cited* 86; 88
Bingham, W. V., *cited* 31
Bishop, O., *cited* 256
Blair, John Lewis, and prediction of school success, 361; *cited* 361
Block, Virginia Lee, 226; *cited* 235
Bobertag, O., *cited* 90
Bogardus, E. S., *cited* 234
Bolton, T. L., 40
Book, William F., and intelligence of city and rural pupils, 410; *cited* 410
Boring, Edwin G., *cited* 267
Bracewell, R. H., and intelligence of city and rural pupils, 410
Bradway, Katherine Preston, and Hoffeditz, E. Louise, *cited* 295
Brainard, Paul P., 228; *cited* 236
Bravais, A., *cited* 63
Brewington, Ann, and clerical aptitude tests, 191; *cited* 191
Bridges, J. W., *cited* 108, 166, 314
Brigham, Carl C., and army tests,

263–264; and intelligence of racial groups, 412, 413; and scores of army groups of different schooling, 423; *cited* 263, 412, 413, 423
Brightness, index of, as test criterion, 162; definition of, 299
Brightness score, use of, 162
Bronner, Augusta F., and Healy-Fernald test group, 177; and tests of special abilities, 239; and special abilities, 391; *cited* 177, 239, 392
Bronner, Augusta F., Healy, William, Lowe, Gladys M., and Schimberg, Myra E., *cited* 31, 182
Brown, Andrew W., and Chicago Non-Verbal Examination, 146; *cited* 164, 168
Brown, Marion A., 222; *cited* 234
Brown, Ralph R., and correlation of retests, 348; *cited* 348
Buckingham, B. R., *cited* 166, 167, 303
Buehler, C., and Hetzer, H., *cited* 165
Burgess silent reading tests, 341–342
Burks, Barbara S., Jensen, Dortha Williams, and Terman, Lewis M., *cited* 348
Burt, Cyril, early correlation study, 72–78; twelve tests of, 72–73; correlation table of tests, 76–77; definition of central factor, 248; and correlation of mental and achievement tests, 400–404; *cited* 155, 400
Byrns, Ruth, and Henmon, V. A. C., *cited* 406

Capacity, not directly measured by tests, 20; tests for analysis of, 169–204; development of specialized tests, 204; correlation of mental tests with other measures of, 353
Capacity, general, subject-matter in tests of, 246
Carter, T. M., and correlation of mental and anatomical age, 311; *cited* 312
Case, A. T., *cited* 235
Cattell, J. McK., and reaction time, 36; and Galton, 37; and early program of tests, 37–38; and 1896 A. P. A. Committee, 39; and Columbia University Tests, 47; and effect of education on intelligence, 426
Cattell, Psyche, and Personal Constant, 296; *cited* 295, 296
Chapman, J. Crosby, and A. Q., 304; *cited* 304
Characteristics, fundamental, of mental tests, 15
Childs, H. G., *cited* 94
Choice of tests, criteria for, 156
Christianson, A. O., *cited* 136
City pupils, intelligence of, 410–412
Clapp-Young Self-Marking Device, and Henmon-Nelson Test, 150; described, 160
Clark, Willis W., *cited* 203
Classification of pupils, 23; vertical, 24; horizontal, 24; into ability groups, 376–384; difficulties of, 377
Classification of tests, 17
Clayton, Blythe, *cited* 270
Clement, John Addison, *cited* 359
Clerical aptitude tests, 190–192
Coefficient, correlation represented by, 48
Coefficient, reliability, 62, 70
Coefficient of correlation, probable error of, 64–65; defined, 337
Coefficient of intelligence, 111; constancy of, 297
Coefficients, tests, hierarchy of, 77
Cole, L. W., and Vincent, Leona, and primary tests, 150; *cited* 165
College success and test scores, 356–360; selection of applicants for, 393
College tests, listed, 167; academic aptitude tests, 192
Colloton, Cecile, *cited* 347
Columbia University, early testing at, 39; description of early tests, 46–51
Colvin, S. S., and definition of intelligence, 248; and correlation of intelligence and college grades, 356; *cited* 167

Comparing results of tests, early procedure of, 57
Completion tests method, development of, 55; in language tests, 272; in non-language tests, 275
Composite personality traits, tests of, 229–233; list, 236
Composite scales, defined, 170; described, 170–182
Composite score, 85; and mental age, 89
Comprehension, Binet's test of, 52
Conduct, relation of intelligence to, 363–369
Conklin, Edmund S., *cited* 235
Constant factor, effect in producing a spurious correlation, 67
Content, a criterion, in the choice of a test, 158
Cornell, Ethel L., and Coxe, Warren W., Performance Ability Scale, 148; *cited* 164
Correction for attenuation, 70
Correlation, general meaning of, 60; between Columbia tests, 48; between average class standing and a number of mental tests, 49; between standing in various college subjects, 50; as method of selecting tests, 250; statistical, 334–344; between mental tests and other measures of achievement, 353–362
Correlation, coefficient of, probable error in, 64; defined, 337
Correlation, partial, 69; as test criterion, 164; and nature-nurture controversy, 403
Correlation, spurious, scatter diagram, to show effect of a constant factor, 67
Correlation, and intercorrelation, relation between, 78
Correlation method, application of 60–84; Spearman's criticism of statistical procedure, 62; correlation study of single tests, 71
Correlation table, 76–77; defined, 335; examples of, 335, 338, 340, 341, 342, 343
Cowdery, K. M., *cited* 235
Coxe, Warren W., 148, 164
Criteria for choice of tests: price, 156; completeness and convenience, 157; adaptation to age, 157; appeal to child, 158; content, 158; length, 158; ease of administration, 159; simplicity of response, 159; ease of scoring, 160; norms, 161; brightness index, 162; tabulation of results, 163; distribution of scores, 163; progression of median, 163; correlation of, 164
Crockett, Alexander C., and motor capacity tests, 198; *cited* 198
Cross-out method, in language tests, 274; in non-language tests, 275
Cunningham, Bess V., *cited* 166

Dearborn, W. F., and correlation of intelligence and school marks, 357; *cited* 31, 165, 166, 357
Dearborn, W. F., Anderson, J. E., and Christianson, A. O., *cited* 136
Decroly, O., and Degand, J., *cited* 90
Definition of tests, 17
Degand, J., *cited* 90
Delinquents, application of mental tests to, 27
Diagnosis, of capacity of pupils, 24
Dickson, Virgil E., and educational use of mental tests, 373; *cited* 373
Differences, inheritance of, 36
Differences, individual, early studies of, 34; individual, 351
Difficulty, of items in a test, 260–270
Directions, simplicity and clearness, a criterion in the choice of a test, 157
Discrimination, tactual, 38, 53; sensory, 72
Distribution of intelligence quotients, 102
Distribution table, 326
Dodd, Stuart C., and International Test, 146
Douglass, Harl R., and Huffaker, C. L., and correlation of I.Q. and E.Q., 307; *cited* 307

Douglass, Harl R., and Spencer, Peter L., *cited* 288
Downey, J. E., and will temperament test, 207–215; *cited* 207, 234
Drake, Raleigh M., and musical memory test, 189; *cited* 189
Dunlap, Knight, and sensory-motor tests, 196–197; *cited* 197
Dykema, Peter W., *cited* 189

Ebbinghaus, H., and completion tests, 55; and concept of intellectual capacity, 248; *cited* 55
Educational guidance, 25; use of tests in, 385–391; and factor analysis, 387
Educational systems, measurement of efficiency, 27
Educational tests, distinguished from mental tests, 18; correlation with intelligence tests, 339
Educational uses of tests, basic facts of, 345–369; prediction of intellectual ability, 345–351; individual differences, 351–353; correlation of mental tests and other measures of achievement, 353–362; mental tests as measures of factors of achievement, 362; relation of intelligence to conduct, 363; group intelligence level and relation to achievement, 370–373; administrative use with individual pupils, 373; determination of time to enter school, 374–376; classification into ability groups, 376–384; selection for special classes, 384; educational guidance, 385–391; adjustment of pupil to work, 391–393; selection of college applicants, 393
Elliott, Richard M., *cited* 184, 236
Engel, Anna M., *cited* 165
Entering school, mental tests as aid in determining right time for, 374–376
Environment, effect of, 28; and mental ability, 396–404
Equation, personal, 34
Error, probable, 35; of sampling, 65; sources of, 279
Errors in measurement or observation, effect on correlation coefficient, 69; types of, 280
Esthesiometer, Binet's, 53; Burt's, 72
Experimentation with tests, early, 34–59; studies of individual differences, 34; American experiments with tests, 38; European, 51; summary, 57
Extroversion, tests of, 224; list, 235

Factor analysis, 17; early theories and development, 78–81; recent work in, 81–84; and aptitude tests, 182; and mechanical aptitude, 185; and special abilities, 194; and profile scales, 201, 203–204; and subject-matter of tests for, 243–246; and educational and vocational guidance, 387; nature of ability, 431–444; concepts and theories of abilities, 433–437; faculty theory, 434; two-factor theory, 435; primary abilities, 437; criticism and evaluation of doctrines, 438–444
Factors, nature of, 82–84; identification of, 244; and common sense, 246; two-factor theory, 435; group factors, 437
Faculty theory, and factor analysis, 434
Fernald, Grace Maxwell, *cited* 171
Fillmore, Eva A., and Iowa Tests for Young Children, 149; *cited* 149, 164
Finality of judgment, in will temperament test, 211
Fischer, Charlotte Rust, *cited* 179
Flory, Charles D., *cited* 153
Footrule method, the, 63
Form-board tests, 186
Foster children, intelligence of, 426–427
Foster, Josephine Curtis, *cited* 33, 149, 165, 314
Franzen, R. H., *cited* 303
Freedom from load, in will temperament test, 208
Freeman, Frank N., and VACO

tests, 153–156; and speed and power in tests, 265; and constancy of I.Q., 291; and acceleration of pupils, 379; *cited* 14, 109, 215, 291, 379, 401, 402, 427
Freeman, Frank N., and Carter, T. M., *cited* 312
Freeman, Frank N., and Flory, Charles D., *cited* 153
Freyd, Max, *cited* 236
Furfey, Paul Hanley, and social reaction tests, 222; *cited* 234

G, 79; definition of, 258–259; and two-factor theory, 436
Galton, Francis, and mental inheritance, 36, 37; influence on mental testing, 56
Galton whistle, the, 37
Garretson, O. K., and Symonds, Percival M., and tests of attitudes, 228; *cited* 236
Garrett, Henry E., *cited* 31, 182
Garrison, S. C., *cited* 347
Gates, Arthur I., and correlation of mental and achievement tests, 354; *cited* 354
General Education Board, 142
General intellectual capacity, early views, 247
General intelligence, existence of, 246–259. *See also* Intelligence
General intelligence level, relation to achievement, 370–373
Geographical groups, differences between, 408–414
Gesell, Arnold, and Pre-School Child Development Scale, 148; *cited* 148, 165
Gesell, Arnold, and Thompson, Helen, *cited* 148
Gifted children, 406
Gilbert, J. A., *cited* 41, 42
Goddard, Henry H., and Vineland tests, 90; revision of Binet scale, 92–94; *cited* 90, 92, 107
Goodenough, Florence L., and drawing test, 146; and intelligence of racial groups, 418; *cited* 164, 418
Goodenough, Florence L., Foster, Josephine C., and Van Wagenen, M. J., and Minnesota Preschool Scale, 149; *cited* 149, 165
Gordon, Hugh, and test scores of children with different schooling, 420
Grade norms, 312
Gray, W. S., and reading disability, 392; *cited* 392
Greene, Edward B., and Michigan Non-Verbal Series, 146; *cited* 164
Griffiths, Nellie L., *cited* 150, 165
Group differences, interpretation of, 421
Group factors, subject matter of tests of, 243–246; and two factor theory, 436; discussion of, 437–438
Group language test, organization of items, 271
Group tests, development of, 2; organization of items of, 271–276; *see also* Point scales
Groups, ability, classification into, 376–384
Growth curves, hypothetical, 292, 293
Growth, mental, and constancy of I.Q., 295
Guidance, educational, 25; vocational, 25; use of tests in, 385–391; and factor analysis, 387
Guilford, J. P., *cited* 82

Haberman, J. V., and Healy-Fernald test group, 177; *cited* 177
Haggerty, M. E., and Delta 1 and Delta 2, and National Intelligence Test, 142; *cited* 165, 166
Haggerty, M. E., Olson, W. C., Wickman, E. K., and conduct tests, 219–220; *cited* 234
Handicapped children, tests for listed, 168
Hardwick, R. S., *cited* 108, 166, 314
Hart, Hornell, *cited* 236
Hartshorne, Hugh, and May, Mark A., and conduct tests, 219; *cited* 219, 220
Hayes, Samuel P., *cited* 168
Healy, William, *cited* 31, 182
Healy, William and Fernald, Grace

Maxwell, test group, 171–178; *cited* 171
Heidbreder, Edna, and extroversion-introversion tests, 224; *cited* 235, 236
Heinis, H., and personal constant, 295; *cited* 295
Henmon, V. A. C., and Nelson, M. J., and Test of Mental Ability, Grades I–III, 150; High School Tests, 152; *cited* 166, 167, 406
Henri, V., 52
Heredity, effect of, 28; studies of, 36; and mental ability, 396–404
Herring, J. P., and revision of Binet test, 112
Hetzer, H., *cited* 165
Hierarchy of intelligences, defined, 78; table of, 77
Hierholzer, Helen M., *cited* 148, 165
High schools, city and rural comparative scores, 411
High school tests, 151; listed, 167; academic aptitude tests, 192
Hilden, A. H., and personal constant, 296; *cited* 296
Hildreth, Gertrude H., 31
Hildreth, Gertrude H., and Griffiths, Nellie L., and Metropolitan Readiness Tests, 150; *cited* 150, 165
Hines, Harlan C., 31
Histogram, of Binet intelligence quotients, 329
Hoffeditz, E. Louise, *cited* 295
Hollingsworth, H. L., *cited* 31
Hollingsworth, Leta S., *cited* 32, 349
Holzinger, Karl J., and bi-factor method, 82, 83; and test scores, 280, 285; and weighted scores, 288; bi-factor theory, 437; *cited* 81, 82, 215, 280, 285, 288, 427
Holzinger, Karl J., and Clayton, Blythe, *cited* 270
Holzinger, Karl J., and Freeman, Frank N., and Burt's regression equation, 401, 402; *cited* 401, 402
Homogeneous grouping, and classification of ability groups, 378
Horizontal classification of pupils, 24; difficulties of, 377
House, S. D., 224; *cited* 235
Huey, Edmund B., *cited* 107
Huffaker, C. L., *cited* 307
Hull, Clark L., 32
Humm, Doncaster G., and Wadsworth, Guy W., Jr., and Temperament Scale, 232; *cited* 236
Hunt, Thelma, *cited* 32, 235
Hunter, Theodore A., *cited* 267
Hunter, W. S., and Sommermeier, Eloise, and mental test scores of Indians, 416; *cited* 416
Hypothetical growth curves, 292, 293

Immigrant groups, intelligence of, 412–414
Impulses, coördination of, in tests of will temperament, 212
Index of brightness, as test criterion, 162; definition of, 299
Indians, scores made by, 416
Individual differences, early studies of, 34; as factor in education, 351–353
Individual pupils, administrative use of mental tests with, 373
Infant and pre-school development schedules, 148–150; list of, 165
Inheritance of differences, study of, 36
Inhibition, capacity for, in tests of will temperament, 211
Institute of Child Welfare, University of Minnesota, 310
Intellectual capacity, and tests, 19; *see* Intelligence
Intelligence, definition for testing, 16–17; test scores measure, 87; factor in military efficiency, 139; and early testing, 169; and school achievement, 205; definitions of, 247–259; measures of relation to achievement, 303–308; constancy and prediction of, 345–351; relation to conduct, 363–369; community level of and school achievement, 370–373; measurement of by tests, 394–396; and heredity and environment, 396–404; vocational group differences in, 404–408; geographical differences in,

408–414; racial differences in, 414–419; effect of schooling on, 419–421; interpretation of group differences, 421–429; of foster children, 426–427; nature of ability, 431–444
Intelligence of Scottish Children, cited 105
Intelligence quotient, in first Stanford revision, 99; developed from Stern's mental quotient, 101; table of distribution of, 102; meaning of, 103; used in Revised Stanford-Binet tests, 105; and army tests results, 127–128; constancy of, 290–300; and constancy of intelligence, 345; constancy of and native ability, 398
Intelligence ratings, distribution in typical army groups, 138
Intelligence test, sample, 3–12; successful because of multiplicity of individual tests, 86; validity of, 254; limitations of, 255; correlation with college achievement, 356–360; interpretation of, 394–430
Intelligence Tests and Their Use, 32
Intelligence and achievement, measures of relation between, 303
Intercorrelation of mental tests, given by Burt, 74, 76; table of, 77; hierarchical arrangement of, 78; as method of selecting, 250
Interests, tests of, 226–228; list, 235–236
Intermediate and upper grades, tests for, 150–151; list of tests for, 166–167
International Business Machine Corporation, electrical scorer, 160
Interpretation of intelligence tests, 394–430; two fundamental problems, 394; differences between vocational groups, 404; differences between geographical groups, 408; differences between racial groups, 414; groups with various amounts of schooling, 419; interpretation of group differences, 421; summary, 429
Interpretation test methods, 276
Interrelationship of mental traits, 28
Introversion, tests of, 224; list, 235
Items of a test, selection and organization, 260–276; difficulty of, 260–270; organization in group language test, 271–275; in non-language test, 275–276

James, William, and definition of intelligence, 249
Jastrow, J., 40
Jennings, H. S., on heredity and environment, 400; *cited* 400
Jensen, Dortha Williams, *cited* 348

Kaufman, H. J., *cited* 165
Kawin, Ethel, *cited* 147
Kelley, Truman L., and factor analysis, 82; and analysis of ability, 245, 259; and primary abilities, 437; *cited* 82
Kent, G. H., and Rosanoff, A. J., and free association tests, 231; *cited* 236
Keys, Noel, *cited* 236
Kindergarten tests, listed, 165
Kingsbury, Forrest A., and Primary Group Intelligence Scale, 149; *cited* 165
Kite, E. S., *cited* 107
Knox, H. A., 147; and test group, 179; *cited* 179
Koerth, Wilhelmine, *cited* 263
Kohlstedt, K. D., *cited* 235, 236
Kohs, S. C., and Block Design Test, 144; and tests of conduct, 221; *cited* 144, 234
Kraepelin, E., and diagnosis of insanity, 55; *cited* 55
Kruger, F., and Spearman, C., 55, 63
Kuhlman, F., revision of Binet-Simon Scales, 93–94; *cited* 32, 93, 107, 295
Kuhlman, F., and Anderson, Rose G., and primary tests, 150; *cited* 165, 166, 167, 168
Kwalwasser, Jacob, and Dykema, Peter W., and musical aptitude tests, 189; *cited* 189

Lafave, Arthur J., 154
Laird, D. A., *cited* 235, 236
Language test, organization of items, 271
Lee, J. Murray, and Symonds, Percival M., *cited* 285
Leland, Bernice, *cited* 201, 202
Length of a test, a criterion of choice, 158; and reliability of, 270–271
Level of intelligence, and relation to achievement, 370–373
Lewerenz, Alfred S., and art tests, 189; *cited* 189
Lincoln, Edward A., and correlation of successive tests, 348; *cited* 348
Lincoln, Edward A., and Shields, Fred J., *cited* 234
Linfert, Harriette-Elise, and Hierholzer, Helen M., and Infant Scale, 148; *cited* 148, 165
Lippmann, Walter, *cited* 126
Loading, in factor analysis, defined, 244
Local norms, use of, 320
Loofbourow, G. C., and Keys, Noel, *cited* 236
Lorge, Irving, and Hollingsworth, Leta S., *cited* 349
Lowe, Gladys M., *cited* 31, 182
Lowell, Francis, *cited* 166
Luria, Max S., and Orleans, Jacob S., and academic aptitude tests, 193; *cited* 193

Maller, Julius B., *cited* 219, 234, 236
Marston, Leslie R., *cited* 234, 235
Matching tests, described, 274
Material, completeness and convenience, a criterion in the choice of a test, 157
Matthews, Ellen, and tests of neurotic tendencies, 225; *cited* 235
May, Mark A., *cited* 219, 220
McAdory Art Test, 190; *cited* 190
McBride, Clifford, 154
McCall, William A., and Multi-mental Scale, 151; *cited* 166
McCarthy, Dorothea, and musical aptitude tests, 188; *cited* 188
Meaning of mental tests, 13
Measurements, significance of, 21
Measures derived from the age scale, 100; between intelligence and achievement, 303
Mechanical aptitude, tests of, 183–187
Meier, Norman C., and art tests, 190; *cited* 190
Melville, Norbert J., *cited* 107
Memory, relationship to estimates of ability, 40; Binet's four tests of, 52
Memory tests, 198; Bolton's, 40
Mental ability, correlation with school success, early studies, 40–51; theory of and factor analysis, 82–84; relative, use of score, 162; specialized, tests of, 239; nature of, 431–444; theories of, 433; two-factor theory, 435; primary abilities, 437; criticism of factor analysis, 438; *see also* Intelligence
Mental age, concept of, 85; and raw score, 277; classification of school pupils by, 377
Mental age method of scaling difficulty of parts of a test, 269
Mental alertness tests, 19
Mental-age scales, *see* Age scales
Mental capacity, and training, 20; and school, marks, 40–41; tests for analysis of, 169–204; *see also* Intelligence
Mental growth, character of, 28; and constancy of I.Q., 291–300
Mental quotient, 101
Mental tests, present status of, 1–29; recent origin, 1; beginning of development, 2; a sample intelligence test, 3–12; reliability and meaning of mental tests, 13; definition and classification of, 17–23; uses of tests, 23; correct answers to sample test, 30–31; uses in the army, 136–140; criteria for choice of, 156–164; *see also* Tests
Merrill, Maud A., *cited* 33, 107
Miles, Walter R., and the Pursuit Meter, 197; *cited* 197
Miller, W. S., and constancy of I.Q., 297; *cited* 167, 297

Mind, British journal, 37, 39, 47
Monroe, Marion, and reading tests, 240; *cited* 240, 392
Monroe, Walter S., and Buckingham, B. R., and Illinois Examination, 152; and Achievement Quotient, 303; *cited* 166, 167, 303
Moral judgments, tests of, 220–222; list of tests, 234
Moss, F. A., Hunt, T., Omwake, K. T., and Ronning, M. M., and social reaction tests, 223; *cited* 235
Motor capacity, Burt's study of, 72; later tests of, 195–198
Motor functions, tests dealing with, 55
Motor impulsion, in tests of will temperament, 210
Motor index, comparison with class standing, 44
Motor tests, Burt's, 72
Movement, speed of, in tests of will temperament, 208
Multiple-choice test method, in language tests, 272; in non-language tests, 275
Murdock, Katherine, and intelligence of racial groups, 417; *cited* 417
Musical aptitude, tests for, 187–189
Myers, Caroline E., and Garry C., and Mental Measure test, 144; *cited* 166
Myers, Garry C., *cited* 166

National Intelligence Test, 3; development of, 142; use of, 150; *cited* 166
National Society for the Study of Education, Thirty-fifth Yearbook, *cited* 380; Twenty-third Yearbook, *cited* 382; Twenty-fourth Yearbook, *cited* 382; Thirty-sixth Yearbook, *cited* 392; Twenty-seventh Yearbook, *cited* 426
Negroes, mental test scores of, 414
Nelson, M. J., *cited* 166, 167
Nemzek, Claude L., and constancy of I.Q., 348; *cited* 348
Neurotic tendencies, tests of, 224–226; list, 235
Newman, Horatio H., Freeman, Frank N., and Holzinger, Karl J., and identical twins studies, 215, 427; *cited* 215, 427
Neymann, C. A., and Kohlstedt, K. D., *cited* 235, 236
Non-language tests, Army Scale Beta, 130–135; later developments, 144–150; list of, 164–165
Norms, criteria in choice of test, 161; and scores, problems of, 277–321; defined, 308; age norms, 309–312; grade norms, 312–313; sex norms, 314–315; race norms, 315–318; social norms, 318–320; use of local norms, 320–321
Nutt, H. W., and rhythm in handwriting, 66; *cited* 66

Occupations, and mental test scores, 404
Odoroff, M. E., and speed *vs.* power in tests, 266; *cited* 266
Oehrn, A., and early tests, 55; *cited* 55
Officers, army, intelligence of, 422
Olson, W. C., 219; *cited* 234
Omwake, K. T., *cited* 235
Opinions, tests of, 226; list, 235
Opposition, resistance to, in tests of will temperament, 211
Organization and selection of items of a test, 260–276
Orleans, Jacob S., *cited* 193
O'Rourke, L. J., and clerical aptitude tests, 191; *cited* 191
Otis, A. S., and army mental tests, 113; and advanced examination scale, 141; and Test of Mental Ability, Alpha, 150; Beta, 150; Test of Mental Ability, Higher Examination, 151; Test of Mental Ability, Gamma, 152; and Index of Brightness, 162; and Classification Test, 153; definition of I.B., 299; and percentile curve, 331; *cited* 166, 167, 331, 344

Parsimony, law of, and factor analysis, 83
Paterson, Donald G., *cited* 164, 181, 191

Paterson, Donald G., and Elliott, Richard M., and test of mechanical aptitude, 184; *cited* 184, 236
Peak, Helen, and Boring, Edwin G., *cited* 267
Pearson, Karl, 56; and products-moment method, 63
Percentile curve, the, 331
Percentile rank, 300
Percentile scores, defined, 269
Perception test, early, 55
Performance tests, described, 147, 148; bibliography of, 164; Knox tests, 179. *See also* Non-language tests
Perrin, F. A. C., and motor ability tests, 195–197; *cited* 195
Personal Constant, 295
Personal equation, the, 34
Personality traits, tests, 205–236; uses of, 29–30; of will temperament, 206–215; of behavior or conduct, 215–220; of moral judgment, 220–222; of social reaction, 222–224; of extroversion-introversion, 224; of neurotic tendencies, 224–226; of attitudes and opinions, 226; of dominant interests, 226–228; of miscellaneous and composite traits, 229–233; comments on, 233; bibliography of, 234–236; and attainment in school, 363; relation to conduct, 367–369
Peterson, Joseph, 32
Philadelphia Mental Ability Test, *cited* 166, 167
Phrenology, and faculty theory, 434
Pillsbury, W. B., and selective elimination of pupils, 421; *cited* 421
Pintner, R., and Non-Language Test, 144; and Mental-Educational Survey Test, 153; and test group, 178–179; and community level of I.Q.'s., 370; *cited* 32, 164, 178, 370
Pintner, Rudolf, and Cunningham, Bess V., *cited* 166
Pintner, Rudolf, and Paterson, Donald G., and test groups, 181; *cited* 164, 181
Point scales, and age scales, 85, 106, 108–111; Yerkes Point Scale, 108–112; Herring Revision, 112–113; U. S. Army mental tests, 113–140; later scales, 141–156; criteria for choice of test, 156–164
Point score, the, 279
Porteus, S. D., and Maze Test, 144; *cited* 130, 144, 164
Power *vs.* speed, in test construction, 262–268
Pressey, S. L., and tests of emotions, 229–231; and prediction of school success, 362; *cited* 362
Pressey, S. L. and L. C., Group Point Scale for Measuring General Intelligence, 142; *cited* 166, 167, 236
Pressey, S. L., and Ralston, R., and intelligence of occupational groups, 405; *cited* 405
Pressey, S. L., and Thomas, J. B., and intelligence of rural children, 410; *cited* 410
Primary abilities, *see* Group factors
Primary grades, tests for, 149–150; list of tests for, 164–165; and kindergarten, tests for, 165–166
Probable error, 35; of coefficient of correlation, 64
Proctor, W. M., and correlation of I.Q. and school marks, 355; and use of tests in educational guidance, 385; *cited* 355, 385
Products-moment method, 63
Professional school, selection of applicants for, 393
Profile scales, defined, 170; described, 198–204
Prophecy law, Spearman's, 271
Proportionality, in intercorrelation of tests, 80, 81
Pursuit Meter, the, 197
Pyle, William Henry, and test groups, 178; and intelligence of racial groups, 418; *cited* 178, 418

Q, a measure of variability, defined, 333
Questionnaire, 36

Race, norms for, 314

Racial groups, differences between, 414–419
Ralston, R., *cited* 405
Rand, Gertrude, and Personal Constant, 297; and distribution of E.Q.s, 308; *cited* 298, 308
Rank method, 63
Rank order, 274
Rating scales, and mental tests, 22
Ratio, accomplishment, 27
Raw coefficient, the, 69
Raw score, the, 277–279
Reaction, sensory-motor, 195; speed and fluidity of, in tests of will temperament, 208; decisiveness of, in tests of will temperament, 209; persistence of, in tests of will temperament, 211
Reaction time, the, 34; 53
Rearrangement, in tests, 275
Recognition test methods, 276
Relative ability, calculating, 162
Relative standing, measures of, 289
Reliability, of tests, 13; factors of, 57; and length of test, 270–271; and test errors, 281
Reliability coefficient, 62; definition, 70; significance, 74
Report of the Committee of the American Psychological Association on the Standardizing of Procedure in Experimental Tests, *cited* 59
Response, simplicity of, a criterion in the choice of a test, 159
Results of tests, how to tabulate, 322–344; tabulating the scores, 322; the distribution table, 326; the percentile curve, 331; correlation, 334
Retardation, 100
"Review of Educational Research," 32
Right-wrong formula, 284
Rogers, Agnes Low, and academic aptitude tests, 192; *cited* 192
Rogers, C. R., *cited* 236
Ronning, M. M., *cited* 235
Rorschach, H., and Ink-Blot test, 231; *cited* 236
Rosanoff, A. J., 231; *cited* 236
Rossolimo, G. J., and profile scales, 199–201; *cited* 199
Ruch, G. M., *cited* 263
Ruch, G. M., and Koerth, Wilhelmine, *cited* 263
Rugg, Harold O., *cited* 269, 344
Rugg, Harold O., and Colloton, Cecile, *cited* 347
Rugg, L. S., *cited* 347
Rural pupils, intelligence of, 410–412

S, and two factor theory, 436
Saam, Theodore, and I.Q. as basis of promotion, 374; *cited* 374
Sample intelligence test, 3–12
Sampling, error of, 65
Scales, Binet, 2, 85; age, 85–106; development of point scales, 108–140; survey of point scales, 141–164
Schimberg, Myra E., *cited* 31, 182
Schmitt, Clara, and Healy-Fernald test group, 176; *cited* 176
Schneck, Matthew R., *cited* 31, 182
Schultz, Richard S., and motor capacity tests, 197; *cited* 197
Scores, tabulation of, 322–326; distribution table of, 326–331; percentile curve of, 331–334; correlation of, 334–344
Scores and norms, problems relating to, 277; mental test scores, 277; the raw score, 277; accuracy of the score, 279; sources of error, 279; treatment of wrong answers, 282; weighting test scores, 286; measures of relative standing, 289; measures of the relation between intelligence and achievement, 303; norms, 308; grade norms, 312; norms for sex, race, and social groups, 314; the use of local norms, 320
Scoring, ease and definiteness, a criterion in the choice of a test, 160
Scrambled Alpha, *see* Army Scale Alpha
Seashore, C. E., and early tests, 43; and musical aptitude tests, 187;

and analysis of musical ability, 240; *cited* 187
Selection and organization of items of a test, 260
Selection for special classes, use of tests in, 384
Sensory discrimination tests, 72, 75, 77, 81
Sensory keenness, Seashore's tests of, 43
Sensory perception, tests for, 53
Sensory tests, recent experimentation in, 195
Sex, norms for, 314
Shields, Fred J., *cited* 234
Shuttleworth, Frank K., *cited* 219
Siceloff, Margaret McAdory, and Others, and art tests, 190; *cited* 190
Simon, T., and early tests, 85, 88; *cited* 86, 88
Simpson, B. R., table of intercorrelations, 81
Skeels, Harold M., and constancy of I.Q., 349; and intelligence of foster children, 427; *cited* 349, 427
Slocombe, C. S., *cited* 348
Social norms, 314
Social reaction tests, 222–224; list of, 234
Sommermeier, Eloise, *cited* 416
Spearman, Charles, and ability analysis, 17; and Seashore discrimination tests, 43; criticism of statistical procedures, 62–71; and factor analysis, 78–84; and "*g*," 79; and analysis of ability, 245; and concept of intellectual capacity, 247, 258; prophecy law of, 271; criticism of Binet theory, 432; two-factor theory of, 435; *cited* 32, 79, 80, 81, 251
Spearman, Charles, and Hart, B., *cited* 79
Spearman, C., and Kruger, 55, 63
Special abilities, tests of, 13; defined, 194; tests of, 194–198; subject-matter of tests of, 239–243
Special classes, uses of tests in selecting children for, 384
Specialized tests of intellectual capacity, development of, 194; selection of subject-matter, 239
Speed *vs.* power, in test construction, 262–268
Spencer, Peter L., *cited* 288
Standard deviation, and constancy of I.Q., 295; and Standard Score, 301
Standard score, and Revised Stanford-Binet tests, 105; calculation of, 301
Standardization, necessity for, 70; definition, 22; and early tests, 57
Standing, comparison with general motor index, 44; correlation with a number of mental tests, 49; correlation between various college subjects, 50; measures of, 289; in comparison with teachers' estimates of ability, 41
Stanford revisions, preliminary investigation, 94; description of first revised scale, 94–98; development of first revised scale, 98–100; intelligence quotient derived from, 100–103; Revised Stanford-Binet tests, 103–106
Stanton, Hazel M., and musical aptitude tests, 188; *cited* 188
Statistical procedure, Spearman's criticism of, 62–71
Stenquist, J. L., Mechanical Aptitude Test, 136, 184; *cited* 184
Stern, William, and early individual psychology, 56; and mental quotient, 101; and definition of intelligence, 248; *cited* 56, 101
Stewart, Frances J., and Brainard, Paul P., and tests of attitudes, 228; *cited* 236
Stoddard, George Dinsmore, and academic aptitude tests, 193; *cited* 193
Strang, Ruth, Brown, Marion A., and Stratton, Dorothy C., and conduct tests, 222; *cited* 234
Stratton, Dorothy C., 222; *cited* 234
Strong, Alice C., and influence of social status on intelligence, 411;

and intelligence of racial groups, 415; *cited* 411, 415
Strong, Edward K., Jr., and tests of attitudes, 228; *cited* 236
Studies of single tests, correlation, 71
Stutsman, Rachel, and Merrill-Palmer Scale of Mental Tests, 147; *cited* 164
Subject-matter of tests, 237–259; selection of, 238; in tests of special capacity, 239; to measure group factors, 243; for general intellectual capacity, and the existence of general intelligence, 246
Sullivan, Elizabeth T., Clark, Willis W., and Tiegs, Ernest W., and profile scales, 203; *cited* 203
Sunne, Dagne, and intelligence of racial groups, 415; *cited* 415
Symonds, Percival M., 228; *cited* 32, 214, 280, 285, 304
Symonds, Percival M., and Block, Virginia Lee, and tests of neurotic tendencies, 226; *cited* 235

T-score, defined, 301
Tabulating results, ease of, as a criterion in choice of a test, 163; methods of, 322–344
Technique and theory of mental tests, 237–321; subject-matter, 237–259; selection and organization of a test, 260–276; problems relating to scores and norms, 277–321
Technique of administration of tests, Burt's, 73
Terman, Lewis M., and first Stanford revision, 94–98; and National Intelligence Test, 142; and Group Test of Mental Ability, 151; and age norms, 310; and sex norms, 314; and constancy of I.Q., 348, 350; and individual differences, 351; and correlation of I.Q. and school marks, 355; and I.Q. necessary for school work, 389; *cited* 33, 107, 167, 347, 348
Terman, L. M., and Childs, H. G., *cited* 94, 107
Terman, L. M., and Merrill, Maud A., and age norms, 310; *cited* 33, 107
Terman, L. M., and Others, *cited* 33, 98, 107, 355, 406
Test construction, principles of, 260–276
Test groups, defined, 170; Healy-Fernald group, 171–178; Pyle group, 178; Pintner group, 178–179; aptitude tests, 182–198; profile tests, 198–204
Tests, mental, reliability and meaning of, 13–17; definition and classification of, 17–23; uses of, 23–30; early experiments with, 34–59; correlation method and, 60–84; age scales, 85–106; early development of point scales, 108–140; in U.S. Army, 113–140; group point scales, 141–164; criteria for choice of, 156–164; for analysis of mental capacity by test groups: composite scales, 170–192; profile scales, 198–204; of personality traits, 205–236; technique and theory of, 237–321; tabulation of results of, 322–344; basic facts of, 345–369; educational uses of, 370–393; interpretation of, 394–430; nature of the ability measured, 431–444
Tests for the analysis of mental capacity, 169–204; Burt's classification of, 72; Columbia University, 46–51; uses of, 23; personality traits, 205–236; profile, 198; correlation studies of, 71
Tetrad differences, 80–81
Theory and technique of mental tests, subject-matter, 237–259; selection and organization of items of a test, 260–276; problems relating to scores and norms, 277–321
Thomas, J. B., *cited* 410
Thompson, Godfrey H., and Burt's regression equation, 402; *cited* 402
Thorndike, E. L., and National Intelligence Test, 142; and Non-Language Test, 144; and correlation of intelligence and college grades, 357; and faculty theory,

434; and intellectual ability, 442; *cited* 357
Thorndike, E. L., and Others, *cited* 33
Thorndike, Robert L., and correlation of retests, 347; *cited* 347
Thurstone, L. L., and factor analysis, 82; and Psychological Examination, 152; and clerical aptitude tests, 191; and academic aptitude tests, 192; and tests of neurotic tendencies, 225; and tests of attitudes, 226; and analysis of ability, 245, 259; and test scoring, 284; and theory of primary abilities, 437; *cited* 4, 33, 82, 155, 168, 191, 235, 284
Thurstone, Thelma Gwinn, 155, 235
Tiegs, Ernest W., *cited* 203
Tomlin, Frank E., and conduct tests, 222; *cited* 234
Toops, Herbert A., and Symonds, Percival M., and A.Q., 304; *cited* 304
Training, effect on behavior, 20
Traits, interrelationship of, 28
Traits, personality, tests of, 205–236
Travis, Lee Edward, and Hunter, Theodore A., *cited* 267
Travis, Lee Edward, and Young, Clarence W., *cited* 268
True correlation, 70
Twins, intelligence of, 215, 427
Two-factor theory, of Spearman, 80, 435

Uses of mental tests, 23; *see also* Educational uses of tests

Validity of intelligence tests, question of, 254
Value of a test, external criteria of, 163
Van Wagenen, M. J., *cited* 149, 165
Variations, among correlation coefficients, causes of, 65
Vector analysis, Thurstone's method, 82
Vernon, Philip E., 227
Vertical classification of pupils, 24; difficulties of, 377
Vincent, Leona, *cited* 165
Vineland Social Maturity Scale, 222
Vocational groups, differences between, 404–408
Vocational selection, 26
Voelker, Paul Frederick, and conduct tests, 215–219; *cited* 234
Volitional perseveration, in tests of will temperament, 212
Voluntary attention tests, 72

Wadsworth, Guy W., Jr., 232; *cited* 236
Washburne, Carleton W., *cited* 382
Watson, Goodwin B., *cited* 235
Watson, John B., *cited* 59
Weber-Fechner Law, 35
Weighting test scores, 286
Wellman, Beth, and constancy of I.Q., 349; effect of education on intelligence, 425; *cited* 349
Wells, Frederic Lyman, *cited* 59
Wenger, M. A., and VACO tests, 153
West, P. V., and test scoring, 284; *cited* 284
Wheeler, L. R., and test scores of children with different schooling, 420; *cited* 420
Whipple, G. M., and Group Tests for Grammar Grades, 141; and National Intelligence Test, 142; *cited* 33, 71
Wickman, E. K., 219–220; *cited* 234
Will temperament, tests of, 206–215; list of, 234
Williams, J. Harold, *cited* 344
Willoughby, Raymond R., and tests of emotions, 231; *cited* 236
Wissler, Clark, and early tests, 39; and Columbia University Tests, 46
Wood, Ben D., and correlation of intelligence and scholarship, 356, 357, 358; *cited* 357, 358
Woodrow, Herbert, and constancy of I.Q., 294; and sex norms, 315; *cited* 33, 294, 315
Woodworth, R. S., and House, S. D., and tests for neurotic tendencies, 224–225; *cited* 235
Woodworth, R. S., and Matthews, Ellen, *cited* 235

Woodworth, R. S., and Wells, Frederic Lyman, *cited* 59
Woolley, Helen Thompson, and Fischer, Charlotte Rust, and test group, 179; *cited* 179
Wundt, Wilhelm, 1, 36

Yerkes, Robert M., and first point scale, 108–112; and army tests, 113; and National Intelligence Test, 142; and race norms, 318; *cited*, 33, 114, 250, 263, 318, 405, 414, 419, 422, 423
Yerkes, Robert M., and Anderson, Helen M., and influence of social status on intelligence, 411; *cited* 411
Yerkes, Robert M., Bridges, James W., and Hardwick, Rose S., first point scale, 108–112; and profile tests, 198–199; and sex norms, 314; *cited* 108, 166, 199, 314
Yerkes, Robert M., and Foster, Josephine Curtis, and sex norms, 314; *cited* 33, 111, 314
Yerkes, R. M., and Watson, John B., 59
Yoakum, Clarence S., and Yerkes, Robert M., on uses of mental tests in army, 137; *cited* 33, 114, 136, 138, 139, 423
Young, Clarence W., *cited* 268

Zyve, D. L., and academic aptitude tests, 192; *cited* 192

DISCHARGED
DISCHARGED
DISCHARGED
DISCHARGED
DISCHARGED
DISCHARGED
DISCHARGED

LB
1131
.F83
1939

39577

STATE TEACHERS COLLEGE
Stevens Point, Wis.

Books may be kept two weeks and may be renewed for one week, except reserved books and magazines. These may be drawn only for overnight use.

A fine of five cents a day will be charged for reserved books not returned on time, and a fine of two cents a day for all other books. No book will be loaned to anyone owing a fine.

LB1131.F83 1939
Mental tests, their history, principles
3 1775 175386 X